Using Item Response Theory to Examine the Psychometric Properties of the Job Content Questionnaire

A Thesis Submitted to the College of
Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the Degree of Master of Education
in the Department of Educational Psychology
and Special Education
University of Saskatchewan

By
Krystal Hachey

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis/dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Educational Psychology and Special Education
University of Saskatchewan
Saskatoon, Saskatchewan S7N 0X1
Canada

Abstract

Hachey, Krystal K., M.Ed. University of Saskatchewan, Saskatoon, DATE. Using Item Response Theory to Examine the Psychometric Properties of the JCQ.

Supervisors: L. Hellsten, B. Noonan, R. Schwier

In the past 30 years, there has been an increase in the number of hours spent in the workforce, and as a result, work stress has been a prominent factor in the increased health problems found in the working population (Briner, 2000). The Job Content Questionnaire (i.e., JCQ) is a self-administered instrument that implements the Demand-Control and Demand-Control-Support models to assess and measure the social and psychological aspects of the work force (Karasek et al., 1998). Thus, the JCQ provides information as to the health of employees. It has been translated and validated in several languages; however each study has only examined the JCQ in terms of Classical Test Theory methods. The current study accumulated validity evidence for the JCQ using Item Response Theory. The results suggested that each of the scales did not contain items that fully measured the latent trait. The analysis also indicated that more items need to be developed. Future research may want to examine other polytomous models, examine males and females separately, and assess the JCQ by the use of Differential Item Functioning (i.e., item bias).

Acknowledgments

This has been a very fulfilling project because I was able to learn not only about the intriguing world of measurement and evaluation, but about myself. Even better, the people along the way provided the best motivation, encouragement, and support. It was these individuals that that I thank, as they are the ones that helped me accomplish this goal.

First and foremost, I want to thank the late Karen Nicholson. She was the one who started me on my wonderful path. She truly was an inspiration of warmth and brightness. For it were those times spent in the lab, that led me to understand and discover the potential and enjoyment of research.

Second, I want to thank Laurie Hellsten for having confidence in me that I could go one step further. It was her that enabled me to stretch my knowledge base in other areas of expertise. With that, she facilitated the quality of my time spent in Saskatoon, as well as, the execution of my thesis.

Third, I want thank Brian Noonan for his words of wisdom and direction. There was not a moment that he was not there to give support and provide information about an area that was new and intriguing.

To Lynn and Gary Hachey, my parents, thank you for your support and guidance. I appreciate all the words of encouragement through my years of schooling.

Finally, I want to thank Adam Arsenault, who has been a constant source of inspiration, guidance and support. He was always there, through it all. For that, I am grateful, for his consistent energy, as it was him who lifted my spirits and always kept me smiling.

Through it all, I made it. However, it was the people around me that helped me get here. So I thank those who have pushed, supported and encouraged. For it was those moments that made me the person I am today.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Name	First Reference
IRT	Item Response Theory	5
CTT	Classical Test Theory	5
DC	Demand Control Model	3
DCS	Demand Control Support Model	3
JCQ	Job Content Questionnaire	3
QES	Quality of Employment Surveys Database	3
DIF	Differential Item Functioning	6
DCQ	Demand-Control Questionnaire	13
OSI	Occupational Stress Index	13
WOM	The Swedish Work Organization Matrix	30
ERI	Effort-Reward Imbalance Model	30
DCSQ	Demand-Control-Support Questionnaire	43
ICC	Item Characteristic Curve	50
ICF	Item Characteristic Function	53
IRC	Item Response Curve	56
IRF	Item Response Function	56
GRM	The Graded Response Model	60
EFA	Exploratory Factor Analysis	8
CFA	Confirmatory Factory Analysis	73
MNSQ	Mean Square	74
ZSTD	Z Standard Deviation	76

CHAPTER I

Introduction

With the increase in prosperity, one would assume that all aspects of the 21st century were on a positive climb. However, this is not the case. Psychologically, individuals have suffered due to work related stress, recognized as a physical ailment. The psychological side of the workforce has been ignored and thus has suffered, which is also a result of modern industrialization (Karasek & Theorell, 1990). Modern Industrialization was marked by the decrease in the quality of goods, the value of services, and having a stronger emphasis on short-term profits. As such, the increase in job participation has resulted in an increase in work stress, which now affects both the work and family environments. Therefore, work stress has developed into an all encompassing malady (Karasek & Theorell, 1990).

In the past 30 years, there has been an increase in the number of hours spent in the workforce, and as result, work stress has been a prominent factor in the increased health problems in the working population (Briner, 2000). With such importance put on work stress and its impact on health, it is imperative to develop a scale that allows companies to measure work stress to incorporate better employee programs (Karasek, 1979).

The work environment, in the area of work stress, is simply referred to as the physical environment in which an individual works. The physical environment can include characteristics of the job (e.g. tasks), broader organizational structure (e.g. history), and even exterior aspects including the extra organizational setting (e.g. labor market). Yet, not all areas of the work environment contribute equally to work stress. Instead, work stress is a combination of the physical and the psychological environment (Briner, 2000).

The psychological environment is encompassed by the physical environment. Briner (2000) has developed two ways in which an individual's psychological environment is created. First, the psychological environment is derived from the individual's interpretation of their environment, and second, from the combination of key work conditions. Physical settings that can influence psychological well being in the work environment can be divided into three groups: (1) Heat, noise and lighting; (2) Nature and social interaction, and (3) The physical environment and physical safety. Therefore, there may be many hidden aspects to a working environment that can decrease psychological well-being (Briner, 2000).

Other areas in the work force that can be influential in the area of work stress include job characteristics, organizational features, and extra-organizational factors. Job characteristics comprise the largest part of an individual's strain and could possibly influence work stress through quantitative and qualitative workloads (i.e., the amount and complexity of the workload; Shaw & Weekley, 1985), task repetitiveness, and role ambiguity (i.e., workers who are lacking in specific information about their role will suffer from stress; Breaugh & Colihan, 1994). Broader organizational factors may be relevant to well-being because of their hierarchical structure (i.e., the way the workers are put into teams), and the culture of the organization (i.e., working hours and how accepting they are of bullying). The last feature of the job force is extra-organizational factors, which can be described on three levels: (1) the individual level, including difficulties outside work (i.e., relationship problems); (2) the community level (e.g. unemployment levels), and (3) the economic level (e.g. industry sector and feelings of job insecurity). For that reason, there are a number of factors that contribute to the area of work stress and the decrease in psychological well-being (Briner, 2000).

The most important issue surrounding work stress is the impact work stress has on health. Evidence suggests that work stress is implicated in the increased rates of absenteeism (Krantz & Lundberg, 2006), high rates of cortisol levels in women (Evolahti, Hultcrantz, & Collins, 2006), a bidirectional effect of Body Mass Index in men (Kivimäki et al., 2006A), high association with depression and chronic pain (Munce et al., 2006) and a relationship with coronary heart disease (Kivimäki et al., 2006B). Furthermore, other negative effects are still being discovered (Briner, 2000). Thus, work stress encompasses every aspect of an individual's life. As a result, it is imperative that there be a way to measure work stress so that work and surrounding environments can be modified and changed to induce worker health.

Model

One of the best known and most widely used models to measure work stress is the Demand Control model (i.e., DC; Karasek et al., 1998). The model examines the interaction between strain and control and the effect of these interactions on the employee. As stated by Karasek (1979) "The job strain model predicts significant variations in mental strain" (p. 8). Johnson and Hall (1988) later added the support aspect to the Demand Control model (i.e., the Demand Control Support model; DCS), after evidence suggesting that supervisor and coworker support could buffer the demands and control of the outcome variables. From these two models,

the Job Content Questionnaire (JCQ) was developed. The model was developed in stages and the core questions were born from three nationally representative samples from the Quality of Employment surveys database (QES). Following a demand for an instrument to examine the DC model and the psychosocial hypothesis, the recommended version was designed (Karasek et al., 1998).

The full recommended JCQ version, has a total of 49 questions (i.e., 5 scales), can be administered in 15 minutes, and measures the mental strain relative to the interaction between Job Demands and Decision Latitude. With the increased relationship of work stress to health variables, it has been translated into a number of languages and reliability and validity estimates have been carried out. On the other hand, the recommended version of the JCQ (i.e., version 1.11; 41 items), has the same number of scales but a smaller number of subscales. The recommended version (i.e., version 1.11 unchanged since 1985) simply differs from the full recommended version because it is the minimum number of items that have been shown to have valid and reliable results (Karasek et al., 1998).

The five scales are (i.e., for both the full recommended and the recommended version 1.11); Decision Latitude (Skill Discretion, Decision Authority, and Skill Utilization), Psychological Demands, Social Support (Supervisor Social Support and Coworker Social Support), Physical Job Demands and Job Insecurity. There is also an umbrella section in which researchers are able to add whatever questions they decide (Karasek et al., 1998). The current study will focus on four of the scales in the recommended version 1.11 (30 items; Decision Latitude, Psychological Demands, Social Support and Job Insecurity).

There have been numerous studies that have either implemented the JCQ to measure work stress or have examined the psychometric properties of the JCQ. However, in all cases, the validity or reliability studies conducted on the JCQ did not use the full or recommended version (Edimansyah, Rusli, Naing, & Mazalisah 2006; Santivirta, 2003; Cheng, Luh, & Guo 2003; Neidhammer, 2001; Brisson, Blanchette, Guimont, Dion, & Vézina, 1998; Storms, Casaer, De Wit, Vandenbergh, & Moens, 2001; Schreurs & Taris, 1998; Kawakami, Kobayashi, Araki, Haratani, & Furiu, 1995; Eum et al., 2006; Sanne, Torp, Mykletunm & Dahl 2005). Thus, there has been inconsistency in the use of the JCQ across the studies.

Although each study examined validity and reliability in terms of the Classical Test Theory (CTT), the JCQ was found to be valid and reliable in a "general" sense. In each case,

some items were incorporated and some items were not. As a result, the evidence is not consistent. Even though the JCQ is the most widely used, and suggested from CTT estimates to be reliable and valid (Briner, 2000), the JCQ may not include items that describe each scale (e.g. Decision Latitude) thoroughly. One way to measure item characteristics as well as person characteristics is through the use of Item Response Theory (IRT; McCarty, 2005). Thus far, no published study has used this theory to accumulate validity and reliability evidence for the JCQ.

IRT, also known as the latent trait model, enables a researcher to examine how an individual would respond to a certain item that measures work stress. As well, it allows the researcher to inspect item characteristics (i.e., difficulty and discrimination), without the constraints (e.g. ability will rise and fall with the difficulty of the items) on item characteristics imposed by the CTT estimates (e.g. Cronbach's alpha or factor analysis). Another limiting factor of CTT is that only one standard error of measurement can be examined, whereas with IRT, standard error for each item and each person can be studied. Thus, there are many advantages in using IRT over CTT (McCarty, 2005).

Purpose

The purpose of the current study was to examine the psychometric properties of the JCQ using IRT. The current study used 30 items and four scales of the recommended JCQ version 1.11. Using Winsteps, a one-parameter IRT program, the following research questions were explored:

1. What is the dimensionality of the JCQ?
    a. What factors and associated items constitute the JCQ?
    b. How well do the items fit each of the resulting subscales?
2. Utilizing Winsteps,
    a. How well does the data fit the model?
    b. How well do the items represent the latent trait (i.e., item quality)?
        i. How does where the range of items fall compare to where the person statistics fall on the latent trait?
        ii. When representing the latent trait, which items overlap and which items are overly spaced?
    c. What are the response probabilities for the polytomous items?
    d. What are the item difficulty placements on the latent trait?

Significance

The current study is significant because there has been a lack of validity and/or reliability studies on the JCQ (i.e., modeled after the DC and DCS model). Previous research has only examined the JCQ using CTT, which cannot investigate each item of the test separately. Unlike CTT, IRT allows researchers to estimate the standard error of measurement at every level of the trait. Another of IRT's advantages is that each item provides information about the latent trait. Furthermore, IRT can allow for the analysis of differential item functioning (DIF), which is also understood as the detection of item bias. Another benefit is that the JCQ tries to measure latent traits (i.e., JCQ assesses psychological measures), and thus is well suited to be evaluated by a latent trait theory (i.e., IRT). Therefore, by using IRT to examine the psychometric properties of the JCQ, the researcher is able to investigate each item of the test, standard error of measurement can be estimated at every level of the test, and DIF can be assessed.

As part of the educational significance of the current study, the advantages of IRT put forth can be used when scales, tests, or questionnaires need to be developed. Adaptive testing is an area that uses IRT and the individual's ability (i.e., the individual taking the test). With adaptive testing, a test can be shortened, as not as many items need to be incorporated without affecting the reliability or validity of the test. If an individual gets an item wrong, then they will get an easy question, and vise versa for getting the question right (i.e., a harder question; Hambleton, 1993). Thus, IRT can be a vital part of test development in the educational field.

Secondary data analysis

Advantages. One of the main advantages of using secondary data analysis is that the data already exits. This can decrease the costs associated with collecting data, as well as, the time involved. Some other advantages include the fact that the size of the sample and its representativeness are already established, and there is less chance for bias (i.e., due to non-response; Sorensen, Sabroe, & Olsen, 1995).

Disadvantages. Some of the disadvantages associated with using secondary data analysis are related to its selection, quality, and the method of how the data was originally collected. Although, using the secondary data means that as a researcher no new data needs to be collected, there may be some areas missing from the sample (e.g. responses; Sorensen, Sabroe, & Olsen, 1995). This is the case for the current study as only 30 items (i.e., only 4 scales) of the recommended version 1.11 were investigated.

Organization of Thesis

Following Chapter 1, which introduces the thesis and examines pertinent definitions, Chapter 2 provides the literature review which is divided up into two sections. The first literature review explains the development and the composition of the DC and DCS model, as well as, the advancement of the JCQ including the scales, validity and reliability studies, translations and studies that have examined the use of the JCQ. The second literature review describes the growth of IRT, the disadvantages of CTT, the IRT model, assumptions, a comparison between the one-,two-, and three-parameter models, polytomous IRT models, Winsteps, and applications of IRT. Chapter 3 provides information as to the methodology of the study. Chapter 3 includes the research design, research questions, sample and data collection, instruments and procedure. Chapter 4 includes both the Exploratory Factor Analysis (i.e., EFA) and IRT analysis results (i.e., using Winsteps), while Chapter 5 concludes with a discussion.

Definitions

Work stress Work stress is defined as the results from an individual continuously having high demands from their work (i.e., high psychological demands) and very little control over what they do (i.e., high decision latitude). Individuals can obtain work stress from a high strain job (Karasek and Theorell, 1990).

Decision Latitude Decision Latitude is the amount of control a worker has over his/her own job (Karasek, Brisson, Kawakami, Houtman, & Bongers, 1998).

Psychological Demands Psychological Demands are the amount of strain an employee gets from his/her job (Karasek et al, 1998). It is also referred to as the workload, and can be defined in terms of time pressure and role conflict (Van der Doef & Maes, 1999).

High strain job A high strain job is the result of individuals who have high demands (i.e., Psychological Demands) and low control (i.e., Decision Latitude; e.g. waitress; Karasek & Theorell, 1990).

Active job An active job is described as having low Psychological Demands and high control (e.g. farmer; Karasek & Theorell, 1990).

Low strain job Low-strain jobs have few Psychological Demands and high levels of control (e.g. repairman; Karasek & Theorell, 1990).

Passive job Passive jobs are the result of having low control and low Psychological Demands (e.g. Sales clerk; Karasek & Theorell, 1990).

Classical Test Theory Classical test theory is a theory that describes test scores by introducing three notions; test score (i.e., observed score), true score, and error score. All together the equation is as follows:

X (observed score) = T (true score) + E (error score)

At any time there are two unknowns in the equation for the examinee, thus, some assumptions must be made. First, true scores and error scores are uncorrelated; second, the average error score in the population is zero, and third; error scores in parallel tests are uncorrelated. In all, Classical Test theory is focused at the test score

	level (Hambleton & Jones, 1993).
Item Response Theory	Hambleton and Jones (1993) state that, "Item response theory is a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test (pg 255)." Therefore, item response theory is focused at the item level (Hambleton & Jones, 1993).
Polytomous items	A polytomous item is one that has more than two score categories (e.g. Likert type format; McCarty, 2005).
Winsteps	A one-parameter item response theory program.

CHAPTER II

Literature Review I: Measuring Work Stress

Introduction

There are a number of models that measure psychological environments and their impact on psychological well-being including General Stress models, the Vitamin model, Affective Events Theory, Psychological Contract, and the Job Demand-Control model. All five models try to explain the relationship between the work environment, the psychological environment, and psychological well-being (Briner, 2000).

General Stress Model

The basic version of the General Stress model depicts the effect of stressors, such as workload and role ambiguity, on a series of "strains", such as mental health and absence. Moreover, there are a number of mediating variables between these connecting strains and stressors, which can moderate the strength of the relationship. These mediating variables include personality, coping and social support. However, the General Stress model is deficient in defining how work characteristics impact well-being and lacks empirical support (Briner, 2000).

Vitamin Model

The Vitamin model was developed by Warr (1994) to suggest a more general approach to the explanation between how mental health is influenced by the key features of jobs and unemployment (Warr, 1994). Warr uses the analogy of the consumption of vitamins and physical health in terms of how some vitamins in high quantities can be harmful to ones health (i.e., A and D), whereas some vitamins have no poor effects on ones health (i.e., C and E). Thus, Warr describes psychological environments as nine environmental vitamins, in which case some, at varying levels, can be detrimental to an individual's well-being (Briner, 2000).

The nine environmental features, which were recognized as important for mental health, include; (1) opportunity for control (e.g. discretion, decision latitude, and autonomy), (2) opportunity for skill use (e.g. skill utilization, required skills), (3) externally generated goals (e.g. job demands, time demands), (4) variety (e.g. variation in job content and location), (5) environmental clarity (e.g. information about the consequences of behavior), (6) availability of money (e.g. income level, amount of pay), (7) physical security (e.g. absence of danger), (8) opportunity for interpersonal contact (e.g. quantity of interaction, absence of isolation), and (9) valued social position (e.g. cultural evaluations of status). As stated by Warr (1994) the only

three environmental vitamins in large dosages that would not produce any ill-effects would be (6) availability of money, (7) physical security, and (9) valued social position (Warr, 1994). However, the relationship between environmental features of Warr's vitamin model and well-being may not always be linear in nature (Briner, 2000).

Affective Events Theory

The Affective Events Theory takes a more specific approach to explaining the environmental features of work and their influence on an individual's psychological well-being. It directly specifies emotions and behavior changes within the environment to describe the changes in the short-term. Thus, it describes events and non-abstract situations that affect individuals, not job characteristics (Briner, 2000).

Psychological Contract

The Psychological Contract was recently popularized by Rousseau (1995) as a theoretical approach to describe how the work environment could affect psychological well-being. The psychological well-being depicts how a worker's beliefs about what they provide to their employer (e.g. effort, commitment) and what they expect in return (e.g. payment, promotion) can influence their psychological environment. If the worker perceives, in any way, that the contract has been broken (i.e., he/she has been providing more effort than required and not getting anything in return), strong negative emotions will be produced. In the long term, these negative emotions can have devastating effects. On the other hand, if the psychological contract is perceived as fair, then psychological well-being will generally improve (Briner, 2000).

Job Demand-Control Model

The most widely tested model is the Job Demand-Control model (i.e., DC) first developed by Karasek (1979). It was initiated to study the effects of cardiovascular disease and worker stress (Theorell, 1996). It shifts the thinking from the individual to the environment, to explain occupational strain (Dollard, 1996). It suggests that the relationship between job demands (i.e., workload) and the well-being of an individual depends on the amount or level of control a worker has. Personal control is regarded as an important aspect to determining health and well-being (Sauter, Hurrell Jr, & Cooper, 1989). This model also proposes that high demands do not always have a negative impact on health (i.e., active job), if a worker has a high level of control (Briner, 2000). Moreover, the process of how the DC model explains the interaction between the psychosocial environment and a worker's psychological well-being, lead

to the initiation of a survey (i.e., Quality of Employment Survey; QES) and to the development of several questionnaires (i.e., Job Content Questionnaire, JCQ; Demand-Control Questionnaire, DCQ; Occupational Stress Index, OSI; Landsbergis & Theorell, 2000) that attempted to measure the construct of work stress (Karasek et al., 1998).

Demand-Control Model

Karasek (1979) introduced the concepts of Job Demands, Job Decision Latitude and mental strain. He proposed that the job strain model was not made up of a single aspect of the working environment, but rather of joint effects from the demands of work situations. Karasek (1979) also reported the interacting effects of Job Demands and Job Decision Latitude. As such, results from Karasek's (1979) work suggested that "active" jobs (i.e., high Decision Latitude and low Psychological Demands) result in the most satisfaction with the least amount of depressive symptoms (Karasek, 1979).

Theorell and Karasek (1990) add to that presented in Karasek's (1979) model of job strain by introducing psychosocial job structure in their book "Healthy Work". Their work bridges the gap between medical science, psychology, sociology, industrial engineering, and economics. The DC model involved job structure, psychological stress, heart disease, and productivity. The authors found evidence relating job structure to psychological stress, which was also found to be related to heart disease. As such, evidence was found that also related job structure to productivity. All areas of interest were affected by age, education and personality. The data was collected from both the United States and Sweden with collaborations from the medical and engineering areas, and social researchers.

Noted by Karasek and Theorell (1990) was the lack of connections between physical ailments due to work-related illness and the psychosocial risks of the environment. They also found that lack of control over an individual's job demands was the major factor relating to the high risk of coronary heart disease. Coronary heart disease symptoms were most common among Swedish working men who described their work as having high Psychological Demands and low ability to make decisions (i.e., the amount of decisions a worker can have measured via a scale; Karasek & Theorell, 1990).

Psychological Demands and Decision Latitude: The Demand-Control Model (DC)

Karasek and Theorell (1990) explain that the DC model is based on three ideas; the demands of work, skill use, and task control, all of which are able to predict a range of health and

behavioral consequences due to work structure (Karasek & Theorell, 1990). There are two hypotheses that have evolved from the development of the DC model; the strain hypothesis and the buffer hypothesis. The strain hypothesis describes the interacting effects of Psychological Demands and Decision Latitude and its affect on job strain, whereas the buffer hypothesis explains how control (i.e., Decision Latitude) mediates the effect of Psychological Demands in the event of job strain (Van der Doef & Maes, 1999).

Strain Hypothesis. Karasek and Theorell (1990) describe the development of the model as consisting of two dimensions; high and low levels of both Psychological Demands and Decision Latitudes. As a result, they developed four types of psychosocial work experiences including high-strain jobs, active jobs, low-strain jobs, and passive jobs. High strain jobs result in the highest chance of employees having a risk of psychological strain and physical illness. Low-strain jobs have few psychological concerns and high levels of control. Finally, passive jobs have low demands and low control, whereas active jobs have both high strain and high control. Figure 1 displays the strain hypothesis and how each psychosocial work characteristic is a function of the interaction between Psychological Demands and Decision Latitude (Karasek & Theorell, 1990).

Active jobs are productive and carry the highest level of performance but do not provide the negative psychological strain that a high strain job does. The amount of work stress depends on the level of control. Individuals who have little or no control over their situation and who are continuously expelling all effort will have high psychological distress. On the other hand, those who have high control and high Psychological Demands will be decidedly productive and have low psychological distress. Results from Karasek (1979) suggest that when there is an increase in both Psychological Demands and Decision Latitude, this leads to increased learning, motivation and development of skill (Karasek & Theorell, 1990).

When a job has high control and high Psychological Demands it is considered an active job, whereas a job with high Psychological Demands and low control is considered a high strain job (see figure 1). Some examples of active jobs include electrical engineers, farmers, and high school teachers. On the other hand, individuals with high strain job include waitresses, nurse's aides, and telephone operators (Karasek et al., 1998).

The other side of Figure 1 is composed of low strain jobs and passive jobs (Karasek and Theorell, 1990). A low strain job has low Psychological Demands and high control and some

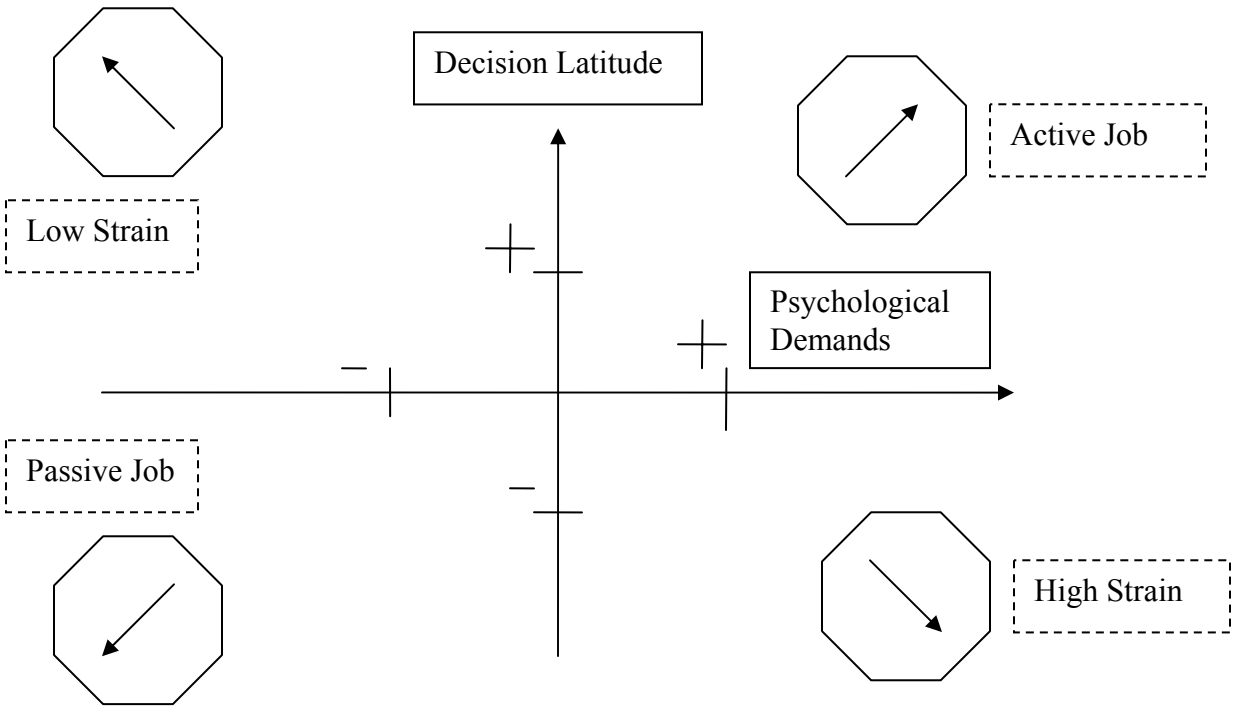


Figure 1. Occupational distribution of psychological demands and decision latitude. Adapted from Karasek et al. (1998).

occupational examples include repairmen or carpenters. A passive job has low psychological demands and low control. An individual in this type of work could be employed as a watchman, billing clerk, or a janitor (Karasek et al., 1998).

Buffer Hypothesis. The buffer hypothesis states that control can cushion the negative effects of high demands, which could also lead to problems with health. As a result, control is a moderating factor between the interacting effects of Psychological Demands and control (i.e., Decision Latitude). Based on the buffer hypothesis, to decrease work stress, job control should be increased without changing the level of Psychological Demands (Van der Doef & Maes, 1999).

Coworker and Supervisor Support: Demand- Control- Support Model (DCS)

Adding to previous work from Karasek's (1979) job strain model, Johnson and Hall (1988) found that the level of social support tended to heighten the effect of job strain. A worker with the lowest rate of social support had higher prevalence rates at each level of strain (Johnson & Hall, 1988). Karasek, Triantis, and Chaudhry (1982) also suggested that social support would moderate job-related stress in the area of physical and mental well-being. However, Karasek et al. (1988) only focused on males in the workplace and had problems with defining different social situations. Thus, their concluding statement was that further tests were needed (Karasek et al., 1982).

As with the DC model, the model proposed by Johnson and Hall (1988) has two hypotheses; the iso-strain hypothesis and the buffer hypothesis. The iso-strain hypothesis and the buffer hypothesis simply have the added dimension of support (i.e., Coworker and Supervisor Social Support).

Iso-Strain Hypothesis. As noted in Figure 2, there is now an added plane to the model; individuals can either be collective workers or isolated workers. Within each group (i.e., collective or isolated), there are the four levels of strain consisting of high and low Decision Latitude and Psychological Demands (Johnson & Hall, 1988). A job that is characterized as having low control, high demands and low support (i.e., isolation) is considered the most harmful working environment and is labeled iso-strain (Van der Doef & Maes, 1999).

Buffer Hypothesis. As with the DC model, the buffer hypothesis states that support, in terms of Coworker and Supervisor Social Support, moderates the effect of high strain. Thus, with

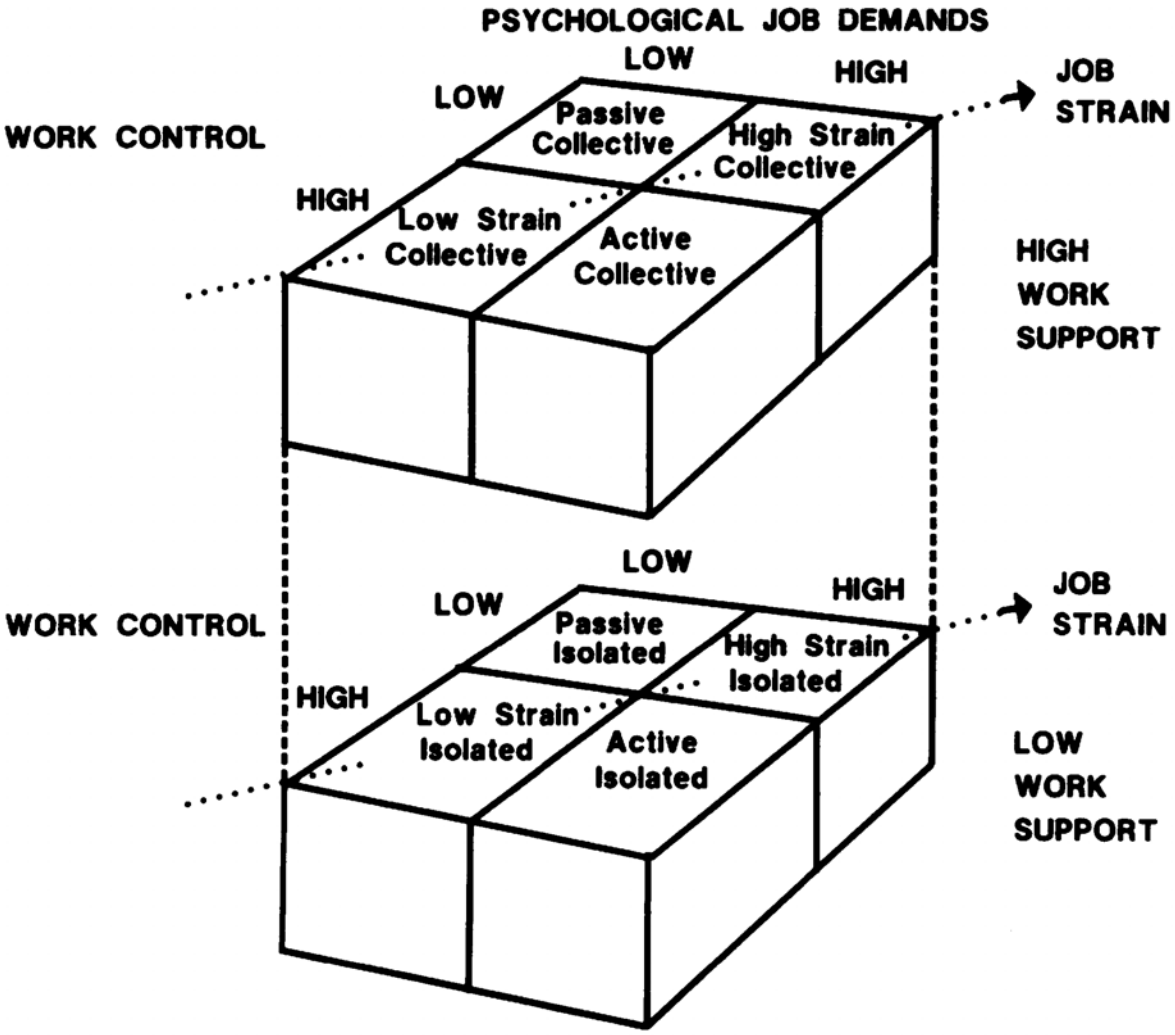


Figure 2. The demand-control-support model. Adapted from Johnson and Hall (1988).

the interacting affects of Psychological Demands and Decision Latitude, Social Support buffers the negative effects on health and well-being (Van der Doef & Maes, 1999).

Measuring Job Demands

There are two ways to measure job demands via the DC or the DCS model, and to clarify the two ways, the definitions suggested by Kristensen (1995) will be used. The two ways include by a "self-rating" and an "average". A "self-rating" (subjective) is obtained by asking each worker questions pertaining to Psychological Demands, Decision Latitude, and Social Support, created by Karasek (1979). The worker is then provided a score that represents their job dimension. On the other hand, an "average" is acquired by either calculating average values from the same (or similar) population or by averaging the score of at least 5 employees to ensure reliability (Kristensen, 1995).

Criticisms

There are two dimensions of criticisms towards the DC and DCS model; theoretical and methodological criticisms. A methodological criticism is that there is a lack of intervention studies. Theoretical criticisms of the DC and DCS model include; (1) that the model is too simple and that more dimensions are needed to explain the interaction between the two environments, (2) Decision Latitude is made up of Skill Discretion and Decision Authority and these two subscales are not always correlated, (3) the model discounts individual differences in terms of coping styles and susceptibility, and (4) the model might be too general to be used on all types of work areas (e.g. education, management) because it was mainly devised for the study of cardiovascular diseases (Kristensen, 1995).

Van der Doef and Maes (1998) conducted research with respect to physical health and (psycho)somatic complaints in conjunction with the buffer and iso-strain hypothesis. Their results, while implementing the DCS model, suggested that depending on the type of approach (physical health versus psycho-somatic complaints), a different hypothesis was more dominant. For (psycho)somatic complaints, the buffer hypothesis was the dominant approach, while the iso-strain hypothesis was the dominant approach for physical health. Thus, this further increases the chance of incorrect outcomes when two sets of approaches are applied to two sets of hypotheses (Van der Doef & Maes, 1998).

The Job Content Questionnaire (JCQ)

The JCQ is a self-administered instrument that implements the DC (and DCS) model to assess and measure the social and psychological aspects of the work force (Karasek et al., 1998). It also has the power to predict musculoskeletal disorders and cardiovascular diseases (Karasek, 1985). Researchers can choose to use the JCQ in several formats, depending on what the needs of the project are. There is the Recommended Format, which has recommended length of 49 questions, and there is a smaller version with limited scales (i.e., version 1.11 with 41 questions; Karasek et al., 1998). There are also additional suggested questions including 22 questions about the work environment, 5 questions about the importance of global economy, 15 questions about technology, and 26 questions about health and well-being outcomes, which also include 18 questions about exhaustion and depression. As mentioned before, there is a final benefit to the JCQ; users are able to construct their own umbrella questions and scale construction equations based on the recommended length of the JCQ format (Karasek, 1985).

There are five scales that make up the recommended version (i.e., version 1.11) including Decision Latitude (Skill Discretion, Decision Authority, and Skill Utilization, Psychological Demands, Social Support (Supervisor Support and Coworker Support), Physical Demands, and Job Insecurity. The purpose of the JCQ is to evaluate stress-related correlates as well as active-passive behavioral correlates. Another benefit to the questionnaire is that it also reviews job security and physical demands (Karasek et al., 1998).

Development of the JCQ

Stage 1: The origin of the JCQ pre-dates 1984 as it was developed in stages. The data was gathered by the University of Michigan Survey Research Center as pooled data to examine job characteristics from three different years (i.e., 1969, 1972, and 1977). The core questions from the JCQ are derived from three nationally representative samples that were born from the Quality of Employment Survey's (QES) database. The survey was not consistent across the three years and asked many different questions in the area of psychosocial job characteristics. A smaller sample of questions was chosen to represent the Job Characteristic Linkage System. Two thirds of those questions, which were similar, were further adapted to create the QES-based JCQ "core". Table 1 presents the core 27 questions from the QES database. In the area of psychosocial job characteristics, it is still the largest nationally representative data set (Karasek et al., 1998).

Table 1

The "Core QES": Number of Items for each Scale

Scale	Core QES JCQ
1. Decision Latitude	
a. Skill Discretion	6
b. Decision Authority	3
c. Skill Underutilization	2[b]
2. Psychological Demands and Mental Workload	
a. General Psychological Demands	4
b. Role Ambiguity	1
3. Social Support	
a. Socioemotional (coworker)	2
b. Instrumental (coworker)	2
c. Socioemotional (supervisor)	2
d. Instrumental (supervisor)	2
e. Hostility (coworker) (new)	
f. Hostility (supervisor) (new)	
4. Physical Demands	
a. General Physical Loading	1
5. Job Insecurity	
a. General Job Insecurity	3
Total Questions	28

Note. QES = Quality of Employment Surveys. Adapted from Karasek et al. (1998)

Stage 2: The initiation in the progression of the JCQ was lead by the U.S. National Heart, Lung and Blood institute. They wished to have a scale developed for the U.S. Framingham Offspring Study, and with that, the same model is in use now. Since the original QES core was not theoretically precise in Psychological Demands and Physical Demands, there was an aim in the development of the JCQ to expand the area of Psychological Physical Demands, Job Security, Social Support, and to aid in discriminant validity. As well, the main goal of the expansion of the JCQ was that it be short, efficient, and self-administered within 15-minutes. Table 2 displays the scales and number of items that make up the full recommended version of the JCQ (i.e., 49 items), which includes scales measuring cognitive workload. The recommended version 1.11 (i.e., 41 items), which has been unchanged since 1985, demonstrates validity evidence and has shown to measure work stress (Karasek et al., 1998). Table 3 displays the scales that make up version 1.11. Appendix A presents a comparison of the QES and the JCQ scales.

A number of goals that motivated the construction of the JCQ included a standard scale reliability assessment, coverage breadth, scale length economy, scale number economy, and specific content interpretability. As well, an important goal was to collect objective data about work environments so there could be prevention oriented aims toward improving the psychological and social working conditions (Karasek et al., 1998).

Scale 1 and Scale 2: Decision Latitude; Psychological Demands. As stated before, Karasek and Theorell (1990) predict that psychological strain occurs when Psychological Demands are high and control is low. Also, what is known as good stress is considered under the active behavior model and it only occurs when psychological strain is high and when control is high (Karasek et al., 1998).

There are two subscales of scale 1 that measure a worker's control over their performance. These scales are components of Decision Latitude and are named Skill Discretion and Decision Authority (there are many more subscales in the full recommended JCQ). Skill Discretion is evaluated by the level of creativity required to accomplish the job, whereas Decision Authority is determined by the workers self-dependence (e.g. autonomy). The first three scales of the JCQ were used to asses the high-demand/low control /low-support model of a workers job strain development (See Table 3; Karasek et al., 1998).

Table 2

Full Recommended JCQ: Number of Items for each Scale

Scale	Full Recommended JCQ
1. Decision Latitude	
a. Skill Discretion	6
b. Decision Authority	3
c. Skill Underutilization	2[b]
d. Work Group Decision Authority (new)	3
e. Formal Authority (new)	2
f. Union/ Representative Influence (new)	3
2. Psychological Demands and Mental Workload	
c. General Psychological Demands	5
d. Role Ambiguity	1
e. Concentration (new)	1
f. Mental Work Disruption (new)	2
3. Social Support	
g. Socioemotional (coworker)	2
h. Instrumental (coworker)	2
i. Socioemotional (supervisor)	2
j. Instrumental (supervisor)	3
k. Hostility (coworker) (new)	1
l. Hostility (supervisor) (new)	1
4. Physical Demands	

Table 2 continued

Full Recommended JCQ: Number of Items for each Scale

Scale	Full Recommended JCQ
b. General Physical Loading	1
c. Isometric Load (new)	2
d. Aerobic Load (new)	2
5. Job Insecurity	
b. General Job Insecurity	4
c. Skill Obsolescence (new)	2
Total Questions	49

Note. JCQ = Job Content Questionnaire

[a] Eight new scales/dimensions and additional items were added to make the Recommended JCQ format. [b] Education was also used in this scale.

Adapted from Karasek et al. (1998)

Table 3

Job Content Questionnaire (JCQ) Recommended Version (Version 1.11, unchanged since 1985; abbreviated wordings)

1a. Skill Discretion
"learn new things ";"repetitive work"; "requires creativity"; "high skill level"; "variety"; "develop own abilities"
1b. Decision Authority
"allows own decisions"; " little decision freedom"; "a lot of say"
1c. Skill Utilization
"education required by job " (also requires education)
1. Decision Latitude
= weighted sum of 1a and 1b
2. Psychological Demands
"work fast"; "work hard"; " no excessive work"; "enough time"; "conflicting demands", "intense concentration"#; "tasks interrupted"#; "hectic job"#; "wait on others"#
3a. Supervisor Social Support
"supervisor concerned"; "supervisor pays attention"; "hostile supervisor"; "helpful

Table 3 continued

Job Content Questionnaire (JCQ) Recommended Version (Version 1.11, unchanged since 1985; abbreviated wordings)

	supervisor"; " supervisor good organizer"
3b. Coworker Social Support	
	"coworkers competent"; "coworkers interested in me"; "hostile coworkers"; "friendly coworkers"; " coworkers work together"; "coworkers helpful"
4. Physical Job Demands	
	"much physical effort"; "lift heavy loads"#; "rapid physical activity"#; "awkward body position"#; "awkward arm positions"#
5. Job Insecurity	
	"steady work"; "job security"; "recent layoff"#; "future layoff"; "career possibilities"#; "skills valuable"#

Note. The symbol # indicates the questions were added in 1985 to create the recommended version.

Adapted from Karasek et al. (1998).

Scale 3: Social Support. Social Support was added to the DC model by Johnson and Hall (1988; i.e., DCS model). Thus, Social Support is a third dimension which can lead to a higher risk of illness in workers. Increased illness can result, if workers control is low, the Psychological Demands are high, and Social Support is low. However, Johnson and Hall (1988) report that there may be gender differences as well as class differences in the interaction of social support in the DC model. Yet, this third dimension demonstrates how important individuals' intrapersonal, as well as interpersonal support, or lack thereof, can affect their psychological well being (Karasek et al., 1998).

Scale 4: Physical Demands. Job strain not only affects the psychological well being of a worker, but can also be demanding on their physical health. If the physical load of a job is high, there will be stress to a workers physical well being. However, high Psychological Demands can also impinge on the physical side of an individual. For example, increased psychological strain can stress the cardiovascular system and musculoskeletal development (Karasek et al., 1998).

Scale 5: Job Insecurity. A final scale added to the JCQ was Job Insecurity (Lohr, 1996). Since changes in the global economy have begun to emerge, there have been greater limitations to maintaining a stable career. Thus, workers must continuously adapt to the changing dynamics of the labor market environment and changes in the labor market have contributed to decreases in job security (Karasek et al., 1998).

Goal and Predictive Validity of the JCQ

The central purpose for the evolution of the JCQ was the ability to gather objective data about an employee's work environment whether it be internal or external. As such, information gathered would be used for the prevention of future deficits in the area of social and psychological working conditions. The JCQ also has strong predictive validity in the area of heart disease and job strain (Karasek et al., 1998).

Questionnaires Developed from the JCQ

Other questionnaires that have been developed from the JCQ include the Swedish Demand-Control Questionnaire (DCQ), the Swedish Work Organization Matrix (WOM), the Whitehall Job Characteristics Questionnaire, and the Effort-Reward Imbalance (ERI) model.

Swedish Demand-Control Questionnaire (DCQ). The DCQ is a shortened and modified version of the JCQ and it was introduced in 1988. As with the JCQ, the DCQ has a scale measuring Decision Latitude, but the scale only contains six questions (i.e., Decision Authority

has two questions and Intellectual Discretion has four questions). The DCQ also has a scale measuring Social Support; however, it is geared more towards the environmental features rather than the objective and instrumental features of the JCQ. The DCQ has adequate internal consistency (i.e., Psychological Demands, $\alpha = 0.75$-0.80; Decision Latitude, α=0.76-0.77; Landsbergis & Theorell, 2000).

Swedish Work Organization Matrix (WOM). The WOM was based off the Level of Living surveys that were administered in Sweden in 1977. Although the WOM is not directly derived from the JCQ there are two items on Job Demands (i.e., Psychological Demands). The WOM is different from the JCQ in that it goes beyond and asks about the selection of supervisors and coworkers, as well as, the planning of vacations. The WOM has internal consistency measures that are higher for the Work-Control scales (i.e., α= 0.75) than the Psychological Job-Demands scale (i.e., α= 0.60; Landsbergis & Theorell, 2000).

Whitehall Job Characteristic Questionnaire. The researchers who carried out the Whitehall study of British civil servants used the JCQ to derive the Whitehall Job Characteristic Questionnaire by adding questions on Decision Authority and changing the format. The format of the questions was changed to a frequency (i.e., 4-point scale that ranged from "often" to "never"). Internal consistency estimates are found to be higher for the Job-Control scale (i.e., α= 0.84) than the Job-Demands scale (i.e., α=0.67; Landsbergis & Theorell, 2000).

Effort-Reward Imbalance Questionnaire (ERI). The ERI expands on the JCQ, DCQ, WOM and Whitehall questionnaire. The Extrinsic effort scale is very similar to Job Demands (or Psychological Demands) of the JCQ (and previous questionnaires), but includes piecework and shift-work. Low reward of the ERI is similar to low social support and is defined as the "esteem reward", including low income and poor job security (e.g. layoffs). The Extrinsic Effort scale and the Rewards scale have good internal consistencies (i.e., Extrinsic reward scale, α= 0.76); Reward scale, α= 0.81-0.82; Landsbergis & Theorell, 2000).

Summary and Critique of the Studies Examining the Demand-Control Model/JCQ

Karasek's (1979) core model and Karasek and Theorell's (1990) DC model has had a popular following since the early 1980's and it is considered to be one of the most influential models measuring work and health. Since its conception, there have been various reconstructions, modifications, and validations, all from the core model. The core model consists of Job Demands and Job Control; while control itself is comprised of two components (i.e.,

Decision Authority and Skill Discretion; Santavirta, 2003). The following section will first review reliability and validity criteria followed by some of the pertinent validity studies dealing with Karasek's (1979) core model. A summary of the English JCQ are in Appendix B and the translations of the JCQ are in Appendix C.

Validity and Reliability Criteria

Reliability. The main concept of reliability is consistency (containing less error). Although CTT explains the theory behind reliability (i.e., broken up into true score and error score), there are different ways to measure reliability (Traub & Rowley, 1991). When examining a test or questionnaire via CTT, the reliability depends on the characteristics of the test, the group of examinees, and the conditions of administration. However, when examining a test or questionnaire via IRT, the group of examinees is independent of the test (Hambleton & Jones, 1993). As such, there is internal consistency (i.e., how well the items fit the construct), test-re test reliability, inter-rater reliability (i.e., the consistency of judges ratings), and alternate forms (Traub & Rowley, 1991). Cronbach's alpha is used to measure internal consistency. Nunnally and Bernstein (1994) state that a reliability coefficient should be marked at 0.70, for instrument construction, as an acceptable coefficient.

Validity. Validity is defined by Messick (1989) as how well the test measures what is it suppose to measure. As such, a valid test wishes to answer the question 'does the instrument provide meaningful scores (Frisbie, 2005)? ' The different types of validity include; Face validity (i.e., does it 'look' valid), content validity (i.e., does the subject matter match the construct), criterion validity (i.e., is it predictive or concurrent), and finally construct validity (e.g. can be; (1) convergent/measure the same thing in another design or, (2) discriminant/ measure the opposite in another design; Cronbach & Meehl, 1955).

Overall, the question we are hoping to answer is; what is the construct validity of the JCQ? Messick (1989) defined two types of threats to construct validity; (1) construct underrepresentation and (2) construct irrelevant variance. The first one signifies that the construct being investigated is not fully covered by the instrument, whilst the second refers to the influence of systematic factors which are not part of the intended construct. As part of the current study, validity will be defined as how the construct(s) are defined within the instrument (i.e., JCQ). However, convergent validity will also be examined (i.e., via dimensionality), as well as concurrent validity.

Validity and Reliability of the English JCQ

There has been a lack of reliability and/or validity studies on the JCQ (i.e., modeled after the DC and DCS model). Karasek (1979) put forth the first study that compared data from the U.S. and Sweden to demonstrate the predictive validity of the Job Strain model. Karasek and Theorell (1990) examined the JCQ, which was developed from the QES surveys, and reported the reliability and validity estimates of the Psychosocial Job Characteristic scales. Karasek et al. (1998) completed a comprehensive study comparing results from four different countries as well as provided their own data to support the JCQ. Lastly, with the most recent validity and reliability estimates, Sale and Kerr (2002) examined the psychometric properties of the Demand Control scale.

As stated before, Karasek (1979) developed a stress management model for job strain, also called the DC model. The model was tested using longitudinal data to determine whether or not, as stated by Karasek (1979), "…workers with jobs that have become more demanding and allow less decision latitude will show more mental strain symptoms at the of the change period than at the beginning (p. 297)". Using the Swedish level of Living Survey, Karasek (1979) asked employees about their job dissatisfaction, life dissatisfaction, pill consumption (e.g. sleeping pill, and the number of sick days) to determine their level of strain. The survey data was analyzed using analysis of variance (i.e., ANOVA) and a multiple regression (Karasek, 1979).

Results indicated that the Job Strain model was a good predictor of variations of mental strain. However, more emphasis was put on what the model could predict then the reliability and validity of the items. Correlations were reiterated from previous research that Karasek presented at a conference (1978; Managing job stress through redesign of work processes) and thus, only the final results were provided. As stated by Karasek (1979), the correlations between discretion and expert ratings were high (i.e., $r = 0.69$, 1968; $r = 0.64$, 1974; $r = 0.78$, 1973; $r = 0.87$, 1971). Karasek (1979) provides no explanation for some of the visible low correlations (although he lists them as high). Hence, Karasek (1979) did not provide the essential validity and reliability estimates, even from a CTT standpoint, to back up the Job Strain model (Karasek, 1979).

Karasek and Theorell (1980) reviewed the statistical validity of the psychosocial work dimensions in the American QES. The three data sets were from 1969, 1972, and 1977 and all participants were between the ages of 18 and 65. Evidence suggests that the scales have sufficient test re-test reliability (i.e., at each year, correlations were above 0.9). Internal

consistency (Cronbach's alpha) results revealed that for men, the coefficients ranged from 0.40 (i.e., Job Insecurity) to 0.83 (i.e., Social Support), whereas for the women the coefficients ranged from 0.36 (i.e., Job Insecurity) to 0.84 (i.e., Social Support). The low coefficients for Job Security should be a cause for concern. Yet, Karasek and Theorell (1990) supply no reason or explanation as to why they are so low or caution use with the Job Insecurity scale. The scale should be above the acceptable level (Nunnally & Bernstein, 1994). Since both males and females received the low coefficients, more attention should be paid to that dimension of the scale.

Karasek et al (1998) provided a comprehensive overview of the JCQ. With evidence from numerous studies, the JCQ was shown to display substantial predictive validity in terms of work related stress (i.e., work stress).There are also high associations with the JCQ and cardiovascular mortality (Landsbergis, Schnall, Warren, Schwartz, & Pickering, 1994), mental strain (Karasek & Theorell, 1990), coronary heart disease (Landsbergis et al., 1994), and musculoskeletal injury Bongers, de Winter, Kompier, & Hildebrandt, 1993).

Karasek et al. (1998) conducted a study to compare means, reliability and validity estimates across six studies, which were carried out in four different countries. They reviewed the studies in the United States (QES), United States (New England Medical Center), Canada-Québec, Canada-Québec (white collar only), Netherlands, and Japan. However, each study did not use the same number of items. Since each study did not incorporate the full version (i.e., recommended full version) or the recommended version 1.11 for the Psychological Demand scale, the five question QES was used to compare. Reliability of the scales was assessed using Cronbach's alpha, concurrent validity was examined by comparing the correlations between scales and subscales, and factor validity was assessed using factor analysis (Karasek, et al., 1998).

Karasek et al. (1998) found that the overall internal consistency coefficients were acceptable (Nunnally & Bernstein, 1994) for both men ($\hat{\alpha} = 0.74$) and women ($\hat{\alpha} = 0.73$). However, the internal consistency results for each of the scales were more inconsistent. For men, the three lowest estimates were found for Decision Latitude (i.e., ranged from $\hat{\alpha} = 0.61$ to 0.71), Psychological Demands (i.e., 5 items that ranged from $\hat{\alpha} = 0.57$ to 0.71), and Job Insecurity (i.e., ranged from $\hat{\alpha} = 0.49$ to 0.74). For women, the three lowest reliability estimates were found for Decision Latitude (i.e., ranged from $\hat{\alpha} = 0.63$ to $\hat{\alpha} = 0.72$), Psychological Demands (i.e., 5 items

that ranged from $\hat{\alpha} = 0.51$ to 0.72), and Job Insecurity (i.e., ranged from $\hat{\alpha} = 0.47$ to $\hat{\alpha} = 0.76$; Karasek et al., 1998). There was quite a range between the six studies and only a couple of the scales had acceptable coefficients at 0.70 (Nunnally & Bernstein, 1994). Yet, both women and men had the same three low scales, which should red flag the reliability of these three scales.

Results from the factor analysis revealed that each of the six studies contained factors that were well defined. In all, the sample from the 1970's shows a clear factor pattern for both men and women. However, Karasek et al. (1998) does not give clear information as to the specifics of each factor analysis. One main cause for concern was the question for "repetitive work". Karasek et al. (1998) found that the item had low and inconsistent loading on the Decision Latitude factor. There were also other items that had low factor loading patterns (i.e., "conflicting demands" and "wait on others"; Karasek et al. 1998). These results show that each of the questions might not be representing the scale that it pertains to, and that further research is needed.

The most recent examination of the psychometric properties of the JCQ was completed by Sale and Kerr (2002). They included a total of 900 employees from hospital resource files in Ontario, Canada. Fourteen core items were incorporated in the study including Decision Latitude (Skill Discretion and Decision Authority) and Psychological Demands. Using Cronbach's alpha, the internal consistency estimates for the scales were $\hat{\alpha} = 0.81$ for Decision Latitude, $\hat{\alpha} = 0.70$ for Psychological Demands, $\hat{\alpha} = 0.77$ for Skill Discretion, and $\hat{\alpha} = 0.63$ for Decision Authority. As stated by Sale and Kerr (2002) correlations for Decision Latitude and Psychological Demands were within the recommended range. Sale and Kerr (2002) suggest that the low reliability is likely due to overlapping items. Therefore, there should be more research completed on the Decision Authority scale as the internal consistency estimates were low (Hensen, 2001).

A confirmatory factor analysis was conducted for each scale (i.e., one and two factor Decision Latitude; one and two factor Psychological Demands), since the exploratory factor analysis had been performed by Karasek (1985). Results indicated that the one factor and two factor Psychological Demands scale were insignificant for goodness to fit index, incremental fit index, and non-normed fit index. Thus, there may be question as to the actual items making up the Psychological Demands scale. According to Sale and Kerr (2000) the results were considered acceptable with enough validity and reliability evidence to employ the scale.

Translations of the JCQ

Malay Version. Edimansyah, Rusli, Naing, and Mazalisah, (2006) carried out a construct validity and reliability study on the Malay version of the JCQ. To ensure face validity, translation (English to Malay from a fluent research officer) and back translation (Malay to English by one of the authors) was completed. However, Edimansyah et al. (2006) only incorporated 21 of the full 49 items and only used the three major scales; Decision Latitude (8 items), Psychological Demands (7 items), and Social Support (6 items; Edimansyah et al., 2006).

Construct validity was investigated using exploratory factor analysis while internal consistencies were examined using Cronbach's alpha. Internal consistency values were acceptable (Nunnally & Bernstein, 1994) for Decision Latitude ($\hat{\alpha} = 0.74$) and the Social Support scale ($\hat{\alpha} = 0.79$), but less than optimal (Nunnally & Bernstein, 1994) for the Psychological Demands scale ($\hat{\alpha} = 0.61$). The results from the exploratory factor analysis revealed that the first factor was associated with the scales of Social Support with factor pattern values ranging from 0.54 to 0.84. The second factor was associated with all areas of Psychological Demands scale with a loading pattern ranging from 0.41 to 0.65. Finally, the third factor was associated with the Decision Latitude scale with factor pattern values ranging from 0.38 to 0.70 (Edimansyah et al. 2006).

The Malay version demonstrated acceptable and satisfactory results for internal consistency (Nunnally & Bernstein, 1994). However, the Psychological Demands scale received the lowest internal consistency (i.e., $\hat{\alpha} = 0.61$) measure. Overall the Malay version was shown to be valid in association with the proper scales. Although there was evidence found for concurrent validity, Edimansyah et al. (2006) had a small sample size (i.e., 50 workers) and a disproportionate amount of females (i.e., 8%). As a sample size less than 100 can produce unstable results (McCarty, 2005). Thus, 50 workers are too small.

Finnish Version. Santavirta (2003) conducted a validity and reliability study on the Finnish version of the Demand Control Questionnaire (Karasek, 1979), which is a shortened version of the JCQ. The purpose of their study was to test the validity and reliability of the Finish version, using only 11 items. The selection of nurses (i.e., 630 nurses) and teachers (i.e., 1028 teachers) was conducted due to the increased susceptibility to develop stress-related illnesses associated within these professions (Santavirta, 2003).

Investigating the Psychological Demands and the Decision Latitude scales, the JCQ was translated into Finnish by two bilingual researchers. Construct validity was addressed, using exploratory and confirmatory factor analysis. Santavirta (2003) examined construct validity on the original JCQ structure (Karasek & Theorell, 1990) and then later divided up Decision Latitude into two latent variables (i.e., Skill Discretion and Decision Authority). Results indicated that the items measuring Skill Discretion could be removed. Internal consistency values were good for the teachers but only mediocre for the nurses with respect to demand. For example, the demand factor for the teachers was $\hat{\alpha} = 0.74$ and for the nurses it was $\hat{\alpha} = 0.66$. In contrast, the Decision Authority factor for the nurses was $\hat{\alpha} = 0.71$ and for the teachers it was $\hat{\alpha} = 0.59$. Thus, the results were the opposite regarding the Decision Authority factor (Santavirta, 2003). In comparison to the acceptable coefficient (Nunnally & Bernstein, 1994), the Decision Authority measure for the teachers and the demand factor for the nurses were both satisfactory.

French Version. Brisson et al. (1998) conducted a validity and reliability study based on 18 items of the JCQ, including the Psychological Demands scale (i.e., 9 items) and the Decision Latitude scale (i.e., 9 items). The purpose of their study was to inspect the internal consistency, disciminant validity, factorial validity, and examine the evidence for a 1-year stability of the JCQ.

Brisson et al. (1998) surveyed a very large population consisting of 8,263 white collar workers from 20 different organizations, in which half of them were women. Since the Decision Latitude scale is made up of two subscales (i.e., Skill Discretion: 6 items; Decision Authority: 3 items), a Likert-type scale was devised with four levels. Two bilingual researchers translated the scales from English to French and the two versions were submitted to two other bilingual researchers to check for consistency (Brisson et al., 1998).

Inter-correlations as well as correlations were examined in the French version of the JCQ. Evidence for internal consistency was found using Cronbach's alpha. Very low correlations were observed between the two scales, which provide indications of some of the independence of the two constructs (i.e., Psychological Demands and Decision Latitude for women r =0.27; for men r = 0.33). Brisson et al. (1998) detected some interesting elements including a positive relationship between men and women and their scores on the two scales. In comparison, Karasek and Theorell (1990) found that there was a negative relationship for women (i.e., Psychological Demands and Decision Latitude for women r = - 0.24). Results showed that overall internal

consistency coefficients for both men and women were at acceptable (Nunnally & Bernstein, 1994) standards (i.e., $\hat{\alpha} = 0$ .74 for men and $\hat{\alpha} = 0.73$ for women; Brisson et al., 1998).

Brisson et al. (1998) conducted a factor analysis of the 18 items and found evidence to suggest that item 32 ("wait on others to complete tasks) may not represent the Psychological Demands scale as originally designed. Although Brisson et al. (1998) tried excluding the item from the analysis; it did not change the results significantly. The French version of the JCQ also demonstrated stability over one year. Over 75% of the workers participated in the study after one year. Pearson correlations coefficients between scores were 0.65 for the Psychological Demands scale and 0.73 for the Decision Latitude scale (Brisson et al., 1998).

Chinese Version. Cheng, Luh, and Guo (2003) evaluated four scales of the JCQ including Job Control, Psychological demands, Supervisor Social Support, and Coworker Social Support. The Chinese JCQ included 22 items and was translated by the first author into Chinese, and then was translated back to English by two bilingual individuals who had no access to the English version of the JCQ (Cheng et al., 2003).

Test-retest reliability, internal consistency, and construct validity were all investigated in the Chinese version of the JCQ. Test re-test reliability outcomes, from three months apart, ranged from satisfactory (Decision Authority, r = 0.64; Psychological Demands, r = 0.62; Coworker Social Support, r = 0.62; Supervisor Social Support, r = 0.36) to moderately reliable (Skill Discretion, r = 0.73). Results indicated that internal consistency for the Psychological Demands scale ($\hat{\alpha} = 0.55$) and for the Decision Latitude scale ($\hat{\alpha} = 0.69$) were below the acceptable level (Nunnally & Bernstein, 1994). When the Chinese version of the JCQ was factor analyzed, the questions "learn new things" and "conflicting work" did not have factor loadings greater than 0.3. In summary, the Chinese version of the JCQ consisted of scales with internal consistency values below the acceptable level and also had items with unsatisfactory factor loadings (Cheng et al., 2003).

Dutch Version. Storms et al. (2001) administered an evaluation of the Dutch version of the JCQ. The Dutch version of the JCQ was based on the JACE study (i.e., A European study of Job Stress, Absenteeism, and Coronary Heart Disease; Houtman et al., 1999), 43 items were applied to the questionnaire, and translations were completed by the second and third authors (Storms et al., 2001). A total of 3,638 workers participated in the study.

The internal consistency estimates of the Dutch version were acceptable (Nunnally & Bernstein, 1994; Decision Latitude, $\hat{\alpha} = 0.78$; Skill Discretion, $\hat{\alpha} = 0.74$, and Decision Authority, $\hat{\alpha} = 0.77$). Findings also showed high correlations between the scales of Decision Authority and Skill Discretion, and the authors suggested combining these scales in future studies (Storms et al., 2001). In all, the Dutch version appears to have adequate reliability.

Storms et al. (2001) also explored discriminant validity and predictive validity. Results showed that their 43-item version, when compared with the General Health Questionnaire, was positively correlated with the Psychological Demands scale and negatively correlated with Decision Authority scale. However, Skill Discretion did not correlate at all (Storms et al., 2001). Therefore, the Dutch version of the JCQ appeared to measure what it was intended to measure in terms of two of the main scales.

Japanese Version. Kawakami, Koboyashi, Araki, Haratani, and Furui (1995) conducted a reliability and validity study on the Japanese version of the JCQ. Participants included 472 men and 108 women from a telecommunications company. The English JCQ was translated from English to Japanese by the first author and an American teacher. The American teacher, who was fluent in Japanese and who was also blind to the English items, conducted the Back translation (i.e., from English to Japanese). Once completed, the back translation was sent to Karasek and items were changed based on his suggestions (Kawakami et al., 1995).

Exploratory factor analysis was conducted to examine the construct validity of the scale. Pearson correlations were completed to assess inter-correlations, while Cronbach's alpha was performed to appraise internal consistency. Results for the exploratory factor analysis revealed that the items for "no conflicting demand" and "non-repetitive work" did not have factor loadings greater than 0.3. Pearson correlations showed that the Decision Latitude score was positively correlated with the Psychological Demands score in both men and women (i.e., men r = 0.31; women r = 0.45). For men, the below acceptable internal consistency coefficients (Nunnally & Bernstein, 1994) were the Skill Discretion scale ($\hat{\alpha} = 0.59$), Decision Authority scale ($\hat{\alpha} = 0.66$), and Psychological Demands scale ($\hat{\alpha} = 0.61$). In all, the Japanese version stated that they had obtained valid and reliable results (Kawakami et al., 1995) yet some of their findings appear questionable.

Korean Version. Eum et al. (2006) translated and back translated the English version of the JCQ into Korean using a bilingual orator. The finished product was examined by Karasek

and some items were corrected according to his feedback. Participants included 338 employees at a university Hospital comprising of nurses, technicians, administrative personnel, and employees of the nutrition department. Eum et al. (2006) incorporated the full recommended 49 items into the Korean version and conducted exploratory factor analysis, Cronbach's alphas and Pearson's correlations. Results provided evidence of validity for Decision Latitude, Psychological Demands, and Social Support (Eum et al., 2006).

Internal consistency coefficients results were split into two groups; nurses and others. Internal consistency values for the nurses group were lower than the other. Despite the non-referenced statement by Eum et al. (2006) that > 0.60 is an acceptable standard for internal consistency, results showed several subscales with below acceptable coefficients (Nunnally & Bernstein, 1994) including Decision Latitude ($\hat{\alpha}$ = 0.66), Skill Discretion ($\hat{\alpha}$ = 0.56), Decision Authority ($\hat{\alpha}$ = 0.57), Psychological Demand ($\hat{\alpha}$ = 0.58), Coworker Social Support ($\hat{\alpha}$ = 0.63), Supervisor Social Support ($\hat{\alpha}$ = 0.69), Job Insecurity ($\hat{\alpha}$ = 0.49), Macro Level Decision Latitude ($\hat{\alpha}$ = 0.57), and Self-Identity through Work ($\hat{\alpha}$ = 0.64). For the "others" group, the subscales with low internal consistency coefficients (Nunnally & Bernstein, 1994) included Decision Authority ($\hat{\alpha}$ = 0.68), Psychological Demands ($\hat{\alpha}$ = 0.61), Coworker Social Support ($\hat{\alpha}$ = 0.68), Job Insecurity ($\hat{\alpha}$ = 0.59), and Macro-Level Decision Latitude ($\hat{\alpha}$ = 0.51; Eum et al., 2006).

Pearson correlations were used for test-retest stability results and factor analysis was used to examine structural validity. In which case, all scales were significant the second time the correlations were run except for the Psychological Demands scale. The Psychological Demands scale dropped by 4 % the second time the test was run. Exploratory factor analysis indicated that each item loaded above 0.3. Therefore according to the theory, Eum et al. (2006) observed correct factor loadings, but very low internal consistency results.

Swedish Version. Sanne et al. (2005) sought to examine the psychometric properties of a shorter version (i.e., Swedish) of the Demand-Control-Support questionnaire (DCSQ). They included 5,227 working individuals in their study, but unfortunately because the participants were from the Hordaland Health Study, types of employment were not known. Specifically, the Social Support items of the DCSQ are slightly different than the JCQ because it is geared more towards the atmosphere of the workplace, whereas the Social Support items of the JCQ are more objective and instrumental in nature (Landsbergis & Theorell, 2000) Since the DCSQ was already in use, no translation occurred (Sanne et al., 2005).

The properties of the scale were examined using principle components analysis and through the calculation of Pearson's correlations and Cronbach's alpha. Results of the principle component analysis suggested that the highest loading items were where they theoretically belonged. Pearson's correlations suggested that the Skill Discretion scale and the Decision Authority scale shared 15 % of the variance. Finally, Cronbach's alpha was calculated to assess internal consistency of the items. The value obtained for the scales and subscales ranged between $\hat{\alpha} = 0.70$ and $\hat{\alpha} = 0.85$. In all, they found that the psychometric properties were "satisfactory" (Sanne et al., 2005).

Summary of Translated Versions of the JCQ

Although the JCQ has been translated into a number of different languages, there is a lack of consistency in the use of the JCQ. First, each study incorporated a different number of items into their version of the JCQ, and second, each study accumulated validity evidence for their version of the JCQ without clarifying what was deemed valid and reliable. Therefore, it may seem that the JCQ has been validated and has been shown to be reliable, but questions remain regarding the consistency and generalizability of the findings.

Each translated version of the JCQ included a different number of items. Item numbers ranged from 11 items (Santavirta, 2003) to 49 items (Eum et al., 2006). As such, it is difficult to compare the reliability and validity estimates when different items were selected for use. The Swedish version (Storms et al., 2001) did use the full recommended version, and the French version (Brisson et al., 1998) added the complete list of items from two subscales (i.e., Psychological Demands and Decision latitude). However, there is a lack of consistency across studies.

Studies Examining the use of the JCQ

With the rise in work stress and the development of the JCQ, there have been a multitude of studies that have implemented the JCQ. De Lange, Taris, Kompier, Houtman, & Bongers (2003) carried out a longitudinal analysis examining the number of studies that incorporated the DC and the DCS model and identified 45 studies. They examined each reference and rated the study as high or low quality based on evaluation criteria. De Lange et al. (2003) evaluated the studies based on design (e.g. incomplete panel), time lags, measures (e.g. psychometric checks on own data), method of analysis (e.g. correlational research), and non-response analysis (e.g. no

check on follow up response). Nevertheless, no study had incorporated IRT to examine validity (de Lange et al., 2003).

Van der Doef and Maes (1999) reviewed 20 years of empirical research and examined 63 samples. They wanted to see if the strain hypothesis was supported and if there was sustainability for the interactions between demands, control, and support. Results suggested that there was more support for the strain hypothesis than the moderating effects of control and support (i.e., buffer hypothesis). In all, no study that was mentioned in Van der Doef and Maes' (1999) review incorporated IRT to examine validity and reliability estimates.

Summary

Although results suggest that the JCQ possess evidence of validity and reliability, the studies that implemented or examined the scale are inconsistent. Even when examining Appendix B, which contains the validity and reliability estimates for the English JCQ, not one study incorporated the same number of items.

Evidence for the validity and reliability of the JCQ was gathered through correlations, Cronbach's alphas and exploratory and confirmatory factor analysis. Yet not one study took the extra step to examine items using IRT so that they could compare items among each other, as well as, study the respondent characteristics. IRT is able to determine exactly where items are not well defined, if some items are overlapping, where more items need to be added, and if the length of the scale should be increased or decreased. Once the scale and items are calibrated, the test itself can be customized (McCarty, 2005). In effect, examining the JCQ with IRT would initiate the consistency among results that is needed for the JCQ.

Literature Review II: Item Response Theory

Introduction

Previous research has only examined the JCQ via Classical Test Theory (CTT) estimates and there is no published study that has assessed the JCQ using IRT. The following literature review will provide an introduction to IRT with respect to the model, polytomous models with focus on graded response models, applications, and the benefits of IRT over CTT. Hence, literature review II will provide a comprehensive overview of IRT.

Historical Issues

IRT, which was first considered the latent trait theory (Weiss, 1983), was very dominant in the measurement world in the 1970's; however, its development can be seen as far back as the 1930-40's (Hambleton & Swaminathan, 1985). IRT, is considered a general statistical theory, but is not directly measurable. IRT can be described as measuring the ability of an individual taking a test (Suen, 1990). The individual's ability can be regarded as a latent trait (Hambleton, 1993). Even though IRT was developed earlier, CTT was at the forefront of measurement and until recently, IRT was forced to take a back seat. However, a rise in the significant limitations of CTT began to surface, and this allowed the reappearance of IRT in the measurement field (Hambleton & Swaminathan, 1985).

Classical Test Theory

Around 1904, CTT, or true-score theory, was established by the pioneering work of Spearman (Ostini & Nering, 2006; Dae-Yeop, 2002) and focused on test level information. CTT is built on measurement concepts which are borrowed from the physical sciences (e.g. error of measurement). Since there is no direct way to measure the phenomena (traits) that interest scientists, CTT was developed to be studied indirectly by measuring other observable variables (Ostini & Nering, 2006). Currently, the text by Lord and Novick (1968) *Statistical Theories of Mental Test Scores* is the most comprehensive piece on CTT (Haertal, 2006).

CTT consist of three concepts; true scores (T), observed scores (X) and error scores (E). Put together, the CTT equation becomes $X = T + E$ (Harris, 1989). CTT involves an additive model (Allen & Yen, 1979). Each individual score is based on only two variables (i.e., true score and observed score) and there are several assumptions underlying this model. The assumptions include; (1) True scores and error scores cannot be correlated; (2) The average population error score is zero; and (3) Error scores on parallel tests are also uncorrelated (Harris, 1989). Since we

cannot determine (T) and (E), we cannot identify the assumptions underlying this model in a straightforward manor, and as such, sometimes this leads to more questions than answers (Allen & Yen, 1979). In all, the underlying notions of CTT have been used to develop many important tests to date, which include determining test length, testing reliability and testing mastery (Harris, 1989).

It was Dr. Frederic Lord who observed the restrictions relating to the application of CTT (Lord & Novick, 1952). Although both true scores and observed scores were dependent on the test, he noted that ability scores were independent of the test. An examinee's ability is determined in terms of the particular test. Consequently, individuals come to a test with prior abilities in relation to what is being measured (i.e., tested ability) on the test. Furthermore, the difficulty of the test can determine whether their true score is high (i.e., easy test) or low (i.e., hard test); whereas their ability score will stay constant no matter what difficulty. As a result, CTT does not take into account previous ability, item discrimination and how test difficulty can influence a true score (Harris, 1989).

Advantages of IRT

CTT is considered a "weak model" because test data can easily meet the assumptions, whereas IRT is considered a "strong model" because the opposite is true. Test data has a difficult time meeting the assumptions of an IRT model because the items must fit the model (Harris, 1989). Thus, if the data fits the model, items and the examinee scores are independent (i.e., local independence; Dodeem, 2004). Moreover, CTT is sample dependant and a linear model, whereas IRT is not. Furthermore, CTT cannot identify any item-ability relationships, whereas IRT can. As a result, there are several benefits to using IRT over CTT (Harris, 1989).

Some other advantages of using IRT over CTT include the use of values from items of a particular group of individuals, and estimating participant ability of only a certain group of items (Hambleton, Swaminathan, & Rogers 1991). CTT has no way of predicting how an individual will respond to a certain item (Hambleton & Swaminathan, 1985). As well, the standard error of measurement for CTT is constant at all levels of the trait, whereas with IRT, researchers have the ability to estimate the standard error of measurement at every level of the trait. The items assessed in an IRT model also have the ability to provide information about the latent trait, whereas with CTT, items cannot be examined separately (Scherbaum, 2006). With the use of

IRT we are able to estimate or predict an individual's ability as well as compare the items across examinees and tests (Hambleton et al., 1991).

Advantages and Disadvantages of CTT

Although CTT has several limitations compared to IRT (i.e., including not being able to separate a participant's ability from the test characteristics, and has restrictions on only being able to compare items within the same or paralleled test; Hambleton & Swaminathan, 1985; not being able to differentiate between one's ability and the difficulty of the test, and the inability to compare individuals and test uniqueness; Hambleton et al., 1991) it has some advantages as well (Hambleton & Swaminathan, 1985). These include that CTT is generally straightforward, mathematically simpler, makes use of smaller sample sizes, and does not require strict guidelines (Harris, 1989).

Two other issues of CTT are the use and definition of reliability, and the fact that we cannot examine each item on a test separately. CTT claims reliability by the use of parallel tests, however in reality, this is very difficult to perform. Since CTT is geared towards the test and not the items in the test, it is difficult to compare how individuals do on different parts of the test. IRT allows for a more comprehensive evaluation of the examinee and the test, as well as a way to predict ability since it captures these features (Hambleton et al., 1991). Finally, with the availability of inexpensive fast personal computers to conduct IRT, individuals are choosing IRT over the computationally simpler CTT (Harvey, 1999).

The Item Response Theory Model

The IRT model is based on two ideas; an individual's ability can be predicted, and the relationship between an individual's ability and the items on a test are described by an item characteristic curve, or ICC (Hambleton et al., 1991). Objectively, IRT aims to look at an observable trait (i.e., an individual's performance) and an unobservable trait (i.e., the individual's ability) on a test, which in turn will be measured by a mathematical function (Hambleton, 1993). The goal of IRT is to model a relationship between a type of variable that cannot be observed, as indicated as an individual's ability, and the probability of an individual obtaining the correct answer (Harris, 1989). Any IRT model may contain one or more parameters underlying the items as well as parameters describing the individual taking the test. Finally, IRT models provide standard error estimates for every individual (Hambleton et al., 1991).

IRT models describe test performance and examinee items, in relation to ability. Response to items can be categorized as dichotomous or continuous and they can be scored either dichotomously or polychotomously. The score categories for items can be ordered or unordered and the IRT model can consist of either one ability, or many. There is flexibility via IRT models which cannot be seen in CTT, (i.e., the ability scale is tied to the items themselves in CTT; Hambleton et al., 1991).

Assumptions of IRT

There are two assumptions underlying the model of IRT. These include unidimensionality and local independence (Hambleton et al., 1991). These assumptions should be met in order to correctly fit data to a model. In instances where the assumptions are not met; the model will be questionable as to whether it is applicable (Hambleton & Swaminathan, 1985).

The assumption of unidimensionality affirms that only one type of ability can be measured by a group of test scores (Hambleton et al., 1991). This is not to say that other abilities cannot affect a test (i.e., levels of motivation and test anxiety), but that there should be a dominant factor which is sufficiently measured by the test (i.e., attachment; Hambleton et al., 1991). This assumption is sometimes difficult to meet because of "other" abilities, including cognitive and personality factors that can influence test performance (Hambleton & Swaminathan, 1985). In all, this assumption specifies the importance of the evaluation through test scores of only one type of ability (Hambleton et al., 1991). Yet in reality, no scale in practice will ever be perfectly unidimensional (Harvey, 1999).

As noted, the assumption of unidimensionality is difficult to meet. Other factors including test motivation, cognitive skills, test anxiety, and test sophistication can influence the amount of abilities brought to a test. As such, these factors can influence the items and the predictability of the main ability in which the researcher may have wanted to study. For that reason, the construct must be well defined and validity evidence must be gathered to ensure that the test measures what it claims to (Hambleton, 1993).

There are a few of approaches which demonstrate that the assumption of unidimensionality has been met. The first approach is to select a model and then fit the items to the chosen model. The second approach is to define the domains in which the researcher is interested in and then choose a model to fit the test. Items are pre-selected and factor analysis (i.e., measuring the variance in unobservable constructs) can be conducted to make sure that the

items fit the dominant ability (Hambleton & Swaminathan, 1985). This is also called confirmatory factor analysis. Conversely, the main idea behind Exploratory Factor Analysis (EFA) is to investigate possible factors. Since it would be difficult to perfectly meet the assumption of unidimensionality, some researchers contend that the main factor must make up at least 20% of the variance (Scherbaum, 2006). Consequently, it is up to the researcher to determine which approach is better in terms of meeting the assumptions of unidimensionality (e.g. Exploratory Factor Analysis: EFA; Hambleton & Swaminathan, 1985).

The second assumption, local independence, states that when abilities influencing the test are held constant, responses to any item are statistically independent. This means that each item is independent of one another (Hambleton et al., 1991). When unidimensionality is met, local independence is usually met as well. Yet, local independence can still be met if unidimensionality has not been satisfied (Scherbaum, 2006). As a result, the complete latent space, which describes the process of inferring from an observed test score, will contain the dominant ability (Hambleton & Swaminathan, 1985).

Local independence specifies that scores on each test item do not present clues to the answers of any other test items. Since both assumptions are quite similar in terms of the latent space, factor analysis methods can also be employed for the assumption of local independence because once unidimensionality is met; local independence is assumed to be met (Hambleton, 1993). Unlike CTT, the data must fit the model chosen; which also infers local independence has been met (Dodeem, 2004).

The item characteristic curve (ICC). The ICC is used in response to the development of logistic curves. However, using these logistic curves has only been a viable option since the advent of the personal computer. The ICC displays the relationship between both the individual's ability level on an item (i.e., the one taking the test) and the probability that the individual will respond correctly to that specific item (Suen, 1990). Figure 3 displays an ICC, with ability placed on the horizontal axis and the probability of displaying the ability on the vertical axis. However, the ICC, or regression, only illustrates a relationship of one construct (i.e., ability). For that reason, if the latent space is multidimensional then it is labeled as an item characteristic function, or ICF (Hambleton & Swaminathan, 1985).

The distribution of the ICC can be defined as the probability that an individual will respond correctly to an item or $P_i(\theta)$. θ denotes the individual's ability level on the trait that will

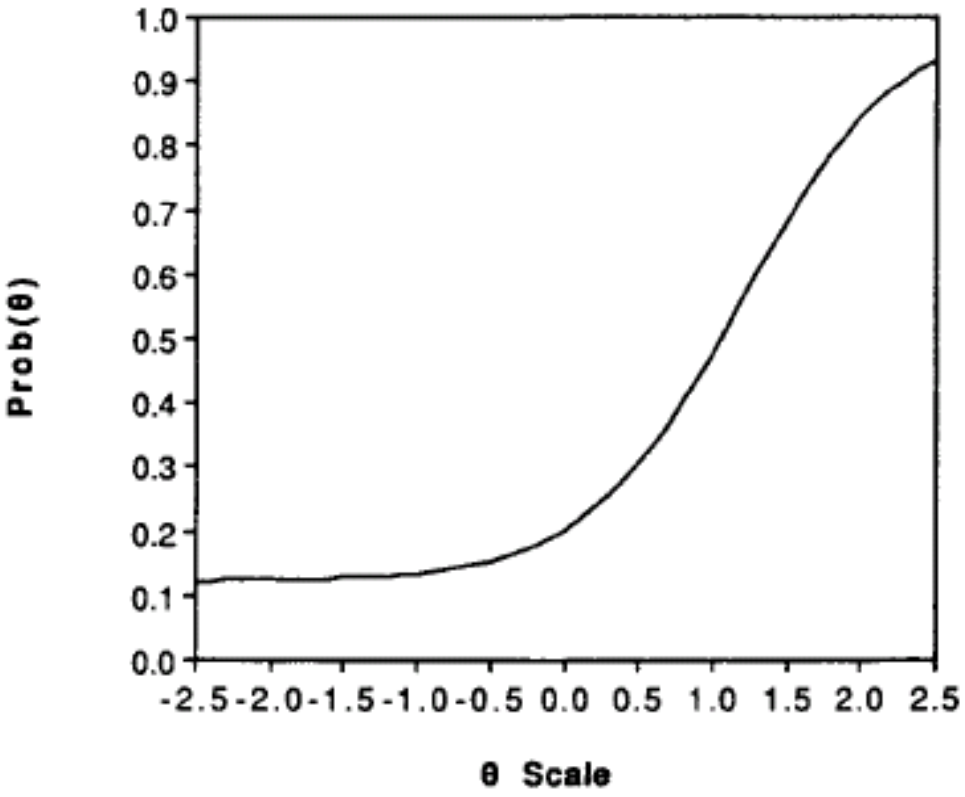


Figure 3. The item characteristic curve (ICC). Adapted from Harris (1989).

be measured, and P_i represents the probability of that individual correctly responding to the item. If we know the association between both the θ and $P_i(\theta)$ for each test item, each item attribute, the ability of each individual, and the measurement error associated with each score, then IRT can be solved mathematically. The ICC is composed of the relationship between θ and $P_i(\theta)$ (Suen, 1990). The non-linear graph that is described as the ICC is made up of connections between the means of the conditional distributions. The equation for these distributions is as follows:

$$f_i(\mu \mid \theta) = Q_i(\theta) \qquad \mu_i = 0 \tag{1.1}$$

That is to say, if the mean of the items is zero at theta, then the probability the individual (i.e., taking the test) will respond incorrectly will be equal to theta (Hambleton et al., 1991). Although confusing, it simply states that the latent trait is the unobservable ability that is defined by θ. The range of an individual's ability is $\in (-\infty, +\infty)$. Yet, one usually does not observe values higher or lower than +/- 3 when scores are scaled with a mean of 0.00 and a standard deviation of 1.00 (Harris, 1989). When scoring a dichotomous item, as the underlying trait increases, so does the probability of a correct response to an item (McCarty, 2005). The θ is also equivalent to the "true score" in CTT (Hambleton et al., 1991). With the ICC, there are also distributions of ability for each examinee. Therefore, both the ICC and the distributions of ability for examinees can be graphed in IRT (Hambleton et al., 1991).

There are almost an infinite number of possible IRT models; however, there are several which are most commonly used. These include the one- parameter logistic models, two-parameter logistic models, and the three-parameter logistic model (Hambleton et al., 1991). Each model has a parameter known as difficulty, also denoted *b*. This can also be seen as the point of inflection on the ability scale (i.e., θ; Harris, 1989). Dichotomous data can be viewed in multiple-choice items, true or false or even short answer (i.e., only looking at the right or wrong answer; Hambleton & Swaminathan, 1985). However, there are more complex models which allow the researcher to carry out their analysis using polytomous or polychotomous data (i.e., opinion survey items or Likert type scales; Harvey, 1999). Overall, IRT allows researcher a way to predict how an individual may respond to a given item on a test from their ability (Harris, 1989).

The Item Response Function (IRF). The IRF is also called the item response curve (IRC) when it is graphically displayed. Ability is displayed along the x- axis, while the probability of

response x_i is on the y-axis. It depicts the probability of a given response to an item. As such, the equation is as follows:

$$P(\theta) \equiv P_i(X = x_i \mid \{\theta\}, \{\delta\}) \tag{1.2}$$

Equation 1.2 portrays the probability that an examinee will have a response $x_{i,}$ if the examinee has the ability level {θ}, and responds to the *i*th item. For dichotomous items the IRC or IRF is synonymous with the ICC, or ICF. This is because the IRF and ICF are equal for dichotomous items. Yet, the term ICC is most commonly used for dichotomous items. However, the IRF, or IRC, and ICC, or ICF, are not the same when discussed within a polytomous item framework (Yen & Fitzpatrick, 2006)

Models of Item Response Theory

The One-Parameter Logistic Model

The one- parameter logistic model, which is similar to the Rasch model[i], is the most extensively used model and carries the equation:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1, 2, \ldots, n \tag{1.3}$$

In equation 1.3, P_i (θ) describes the probability that an individual, with ability θ, will respond correctly to a given item of *i* (Hambleton et al., 1991). The ability of the individual, also denoted as (θ), is continuous but varies in accordance to the distinctiveness of each item of the test. The *e* is the transcendental number whose value is 2.718 (Harris, 1989), while b_i is the difficulty parameter of the item, *i* describes which item, and *n* expresses the number of items in that scale. The more difficult the item, the more difficult it is for the individual taking the test to answer it correctly (Hambleton et al., 1991). Items found on a test that have low values of *b* are considered to be easy, and vice versa for higher level values of *b*, which supports the theoretical definition of *b* (Harris, 1989). It is assumed that in a one-parameter model, that item difficulty is the only thing that influences an individual's performance. For a dichotomous item, difficulty is defined as easy and hard (i.e., there are only two response choices). However, for a polytomous item (i.e., more than two response categories) difficulty is defined a little different as there are more than one choice (Hambleton et al., 1991).

Another interesting aspect of this model is that the lower asymptote (i.e., straight line to which the ICC approaches zero), suggests that an individual with a very low ability will have a near zero probability of correctly answering the item. Therefore, the one-parameter model does

not account for the fact that individuals with low ability may in fact guess (e.g. in multiple-choice). Nonetheless, model selection depends on how the data is scored (e.g. multiple-choice versus Likert scale; Hambleton et al., 1991).

The Two-Parameter Logistic Model

The two-parameter logistic model is a generalization of the one-parameter model and holds the equation:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n \qquad (1.4)$$

The added features in the two-parameter model are included in equation 1.4 as the a_i and D. The added factor of D is a scaling factor, which in turn moves the curve as close as possible to the normal ogive curve (i.e., a normal distribution; Hambleton et al., 1991). Lord (1953) was the first to propose such a model from the normal ogive curve; however, it was Birnbaum (1968) who suggested the two-parameter model take the form of a logistic model rather than a normal ogive curve. Logistic curves are more convenient than trying to use normal ogive curves (Hambleton & Swaminathan, 1985).

This model indicates that guessing cannot occur. This is indicative of the added feature of a_i, which signifies a positive relationship between performance and ability. Meaning, as the probability of guessing a correct answer decreases, ability decreases as well (Hambleton & Swaminathan, 1985). The level of difficulty for each item is $\in (-\infty, +\infty)$. Yet, in practice item difficulties tend to be smaller than |2.0| (Harris, 1989). This is also because items with steeper slopes make it easier to separate examinees than items with gradual slopes. Therefore, each item has the capability of discriminating (i.e., a) at each ability level θ (Hambleton, 1993).

The Three-Parameter Logistic Model

The three- parameter logistic model includes variables from the one- and two parameter logistic models and the equation is written as follows:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n \qquad (1.5)$$

The three-parameter logistic model takes into account lower ability levels and for the possibility of guessing. The added variable c_i, shown in equation 1.5, is named the *pseudo-chance-level parameter* and can be used for multiple-choice test items for which guessing can be a factor (Hambleton et al., 1991). The c_i parameter, also defined as a lower asymptote, is defined by

Hambleton and Saminathan (1985) "as the probability that individuals with lower ability correctly answering an item on a test". It can range from 0.0 to 1.0, but is usually smaller than 0.3 (Hambleton & Swaminathan, 1985). The added parameter gives a greater chance that even individuals with lower abilities may be able to answer correctly to either moderate or hard items on a test (Harris, 1989). If *c* is equal to zero, implying that the probability of a low ability individual answering an item correctly is near zero, we effectively have a two-parameter logistic model (Harris, 1989). To that extent, the three-parameter logistic model includes another characteristic for predicting test ability which the other two models were lacking (Hambleton & Swaminathan, 1985). Nevertheless, a two-parameter model is sometimes preferred since it can be difficult to estimate this lower asymptote denoted as *c* (Harris, 1989).

The difference between one - parameter, and two - and three- logistic models is that the latter have an item discriminatory variable denoted as *a*, which is also the curve of the ICC after the inflection point. The higher this value, the more the item differentiates between individuals taking the test (Harris, 1989).

Other models used in IRT include the four – parameter logistic model, nominal response model, and the graded response model. A four-parameter logistic model takes into account high-ability examinees that may be careless and actually have more information than assumed by the writer of the test. For that reason, the equation is written as follows:

$$P_i(\theta) = c_i(\gamma_i - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n \tag{1.6}$$

This model introduced a new term into the three –parameter logistic model, y_i, as shown in equation 1.6. It assumes that the ICC may have an upper asymptote less than one. Yet, there are not as many practical benefits observed to using this model as the one-, two, or three- logistic models (Hambleton, 1993).

Polytomous IRT Models

Polytomous IRT models are used when there are more than two response categories to score an item. Polytomous IRT models can fit into three categories; graded response models, sequential models, and partial credit models (van der Ark, 2001). Polytomous IRT model formats are readily used mainly because they provide more information and are more reliable than dichotomously scored items (Embretson & Reise, 2000). A nominal response model can be employed when test items are scored on a rating scale, such that they are multi-dichotomous

(Hambleton & Swaminathan, 1985). On the other hand, a graded response model incorporates the usual assumptions with the added classification of an ordering type scale (i.e., Likert type). The data that can be used in these types of tests include multi-chotomous and continuous data. Multi-chotomous data arises when weights can be attached to the listed choices (i.e., Likert scales), whereas continuous data can be observed on a continuous rating scale (Hambleton et al., 1991).

Samejima's Graded Response Models

The graded response model (GRM) is appropriate when response options for items are sequentially ordered (e.g. 0, never; 1, sometimes; and 2, always; Ostini & Nering, 2006) and are most appropriate for attitude and personality measures, but not limited to these domains (Reise & Yu, 1990). To score the item, one would usually look at the probability of obtaining above expected categories (i.e., 2), rather than lower category scores (i.e., 0 or 1; McCarty, 2005). The GRM is also a polytomous IRT model that is a materialization of Thurstone's method (i.e., 1947; successive intervals). The equation for the GRM is

$$P_{ig} = \frac{e^{a_i(\theta - b_{ig})}}{1 + e^{a_i(\theta - b_{ig})}} \tag{1.7}$$

In equation 1.7, the a_i is the item discrimination; the b_{ig} is the boundary locator parameter (i.e., category), and P_{ig} is the cumulative boundary function. However, P_{ig} can be modeled by any appropriated mathematical function (e.g. 2- parameter normal ogive dichotomous model; Ostini & Nering, 2006). As such, the GRM is used when there are three or more graded or ordered scoring categories (Thissen & Wainer, 2001).

Since model selection depends on how the data is scored (e.g. multiple-choice versus Likert scale; Hambleton et al., 1991), the current study will focus on a one-parameter IRT model because the data has been polytomously scored. A graded response model will be implemented because the data are sequentially ordered (e.g. 0, never; 1, sometimes; and 2, always; Ostini & Nering, 2006).

Partial Credit Models

The partial credit model follows the historical development of the rating scale model. The purpose behind this model is that with each part of the problem (e.g. an examinee is answering an item) the examinee has a chance to earn some credit. This is also called item steps. Since it is a flexible model, categories can vary in number as well as items (Ostini & Nering, 2006).

Sequential Models

Sequential models are based on the assumption of monotonocity (i.e., non decreasing ability θ). However, there is a limited number of programs that will actually support this type of polytomous IRT model and usually several programs must be utilized. As such, a researcher must be practical and choose the program with the most desirable traits (van der Ark, 2001)

There are several reasons why researchers would choose to use polytomous data over dichotomous data. First, to obtain the same reliability as dichotomous data, fewer polytomous items are needed. Second, some traits can be more easily measured when using a rating scale, and finally, some categories of variables are expressed more readily on an ordinal scale (van der Ark, 2001).

There are several motives for employing polytomous IRT models. First, that polytomous items exist within psychological measurement (e.g. personality, social, etc). Prior to IRT, one had to utilize Thurstone (1947) or Likert type scaling for polytomous data; however, there were limitations to these approaches. For Thurstone (1947) scaling to work appropriately, one must assume that the trait is normally distributed or that items are chosen so that a bell curve occurs. On the other hand, Likert scaling must assume a linear relationship between the response likelihood and the latent trait. Thus, using polytomous IRT models are considered more appropriate when faced with polytomous items (Ostini & Nering, 2006).

In terms of psychometric issues, there tends to be more positive attributes to using polytomous items over dichotomous items. Polytomous items cover a wider area of the trait continuum; have a greater number of response categories, which in turn, can provide more information about the trait level. As well, polytomous items provide more detailed investigative information about the items and the participants responding to the test, whereas reducing a multiple category test to dichotomous items can leave out systematic measurement information. Some practical advantages include decreased test time and cost, as well as an increased positive effect on participant motivation to complete a test with fewer items (Ostini & Rening, 2006). Therefore, the use of polytomous IRT models is beneficial in terms of their usage and implementation.

Estimation Procedures

An IRT model contains ability as well as item parameters; however the values for these are unknown. Only the test items are known, while the ability and item parameters are unknown,

and therefore must be estimated (Hambleton et al., 1991). These can be done by procedures including Maximum Likelihood, Marginal Maximum Likelihood, and Bayesian Estimation Procedures (See e.g. in Hambleton et al. 1991).

Since there are an infinite number of logistic models that can be configured, and the parameters are difficult mathematically to produce, there are computer programs to complete such IRT models; specifically one-, two-, and three- parameter logistic models.

Winsteps

One such program that uses IRT is Winsteps. It is a Windows based software program that applies the Rasch model (i.e., one-parameter model) for persons and items. Winsteps is usually applied in the area of educational testing, rating scale analysis, and attitude surveys. Winsteps begins with a central estimate for each person measure, item calibration and response-structure calibration. This program sets some ranges on *a, b*, and θ parameters to ensure that the means and standard deviations are 1 and zero correspondingly. It scales the θ parameter, which sets a base for the *a*'s and the *b*'s, however the *c*'s are scale free. The program produces the output in the forms of graphs, plots and tables. More information on Winsteps is provided in the Winsteps professional manual (Linacre, 2006).

Limitations

Even though IRT tends to overcome many of the shortcomings of CTT, it has its limitations. First, IRT is very complex and item parameter estimation problems are apt to arise in practice. Another problem can take place with model fit and dimensionality. Although one-parameter logistic models are more straightforward to apply because they only have one parameter, they are usually questionable when it comes to data fitting because the assumptions are so restricting compared to CTT. Some other limitations of IRT models have to do with the strong assumptions following the theory, how the theory is not robust to violations, and all cases of omitted responses. Moreover, IRT requires a larger sample size (i.e., sample size for CTT, 200-500; sample size for IRT, depends on model, but are usually larger than 500; Hambleton, 1993). However, given its few shortcomings, IRT is typically chosen over CTT methods for measurement practices (Harris, 1989).

IRT is based on a mathematical model. The complexity of this model requires the use of a computer to carry out the analysis (in some cases the researcher may be able to mathematically derive the equations if the necessary tools are available). Also, the ICC scale is arbitrary and has

no direct relationship to the observable data. Scales cannot be compared because the metrics in the computer programs may be used differently. Thus, sometimes just the complexity itself can deter individuals from using IRT (Loyd, 1988).

As noted, one significant limitation of the IRT model is sample size. In some cases, the sample size required for an IRT model (i.e., starting at 150 participants) is just too large for organizational research. Some more complex models can involve even larger sample sizes, which again can make it very problematic for researchers to carry out the IRT design. As a result, individuals may be more apt to choose a design like CTT because they can carry out designs with a smaller sample sizes (Scherbaum, 2006).

Another limitation of IRT is its lack of robustness to violations towards assumptions. There have been suggestions that IRT cannot be applied to all situations, specifically situations where guessing may occur. However, with a three parameter model, the chance of guessing can be controlled for (i.e., multiple-choice; De-Ayala, Plake, & Impara, 2001). Practitioners also face the problem of making sure that the assumptions are met. It may take more time to meet assumptions since IRT is not following an observable trait (Hambleton, 1993).

Applications of Item Response Theory

Once validity evidence is collected for the JCQ using IRT, there are some promising future applications of IRT to the JCQ domain. These future applications include additional instrument development, examination of differential item functioning (i.e., DIF) and adaptive testing (Hambleton, 1993).

Test Development

As listed by Hambleton (1993) the steps of test development for IRT are as follows: (1) Preparation of test specifications, (2) Preparation of the item pool, (3) Field testing the items (i.e., data collection and analysis), (4) Selection of test items, (5) Compilation of norms (for norm-referenced tests; data collection and analysis including scaling), (6) Specification of cutoff scores (for criterion-referenced tests), (7) Reliability studies, and (8) Validity studies.

The purpose of IRT is to predict the ability or traits measured by the test. Thus, a researcher starts out with a set of items responses from a set of individuals who have responded to a test (Hambleton, 1993). Selection of the items is based on the purpose of the test (Harris, 1989). Then, an IRT model is selected and fitted to the response data, while an application of data items is being collected. Once completed, the ability scale is set in accordance with the IRT

model and the test data. The resulting ICC will have the desired mathematical form (Hambleton, 1993). In all, it is a very complex process to create the desired mathematical form from test data.

Test development in the IRT setting includes data banking, tailored testing, test equating and pattern analysis. Data banking involves the calibration of test items to the same scale. There is also the inclusion of sub-scaled items. Test equating entails the researcher to choose the most discriminating items to most accurately predict the ability of a specific examinee with the most effectiveness. On the other hand, pattern analysis deals with individuals that do not fit the pattern by looking at unusual pattern responses. IRT test development, although involves much complexity, embraces promise in the area of measurement (Loyd, 1988).

When conducting an IRT test, the appropriate test length and/or sample size must be considered. For instance, having several model parameters usually means that the sample size needs to increase (i.e., the choice of the IRT model). Second, how the model is applied is imperative. Third, the distribution of the ability (i.e., heterogeneous versus homogenous) must be assessed. Fourth, what is being estimated (i.e., items or parameters) must be clear, and finally, the importance of the application of the IRT model (Hambleton, 1993).

Adaptive Testing

Adaptive testing is also considered tailored testing. The purpose of this type of test, which is usually processed and completed on a computer, is to match the difficulty of the test to the ability of the individual taking the test. Hence, if the individual answers a question correctly the next question would have a higher difficulty level. On the other hand, if the individual answers a question incorrectly the next question would be easier. Surprisingly, the researcher will only need about 50% of the items from a conventional test to produce the same quality. The test can be much shorter without any drop in measurement quality (Hambleton, 1993).

Item Bias

Item bias can be a huge issue if individuals respond differently to items based on ethnicity, culture, sex etc. Therefore, there needs to be a way for researchers to be aware of item bias and to have a means of removing it. For example, if ethnicity is an issue, ICC's must be plotted for both groups and then compared to see if they do (i.e., unbiased) or do not (i.e., biased) match. It is a relatively simple procedure. The IRT model is chosen between one-, two- or three-parameter logistic models and then the latent trait is chosen. In all, IRT permits the researcher to

compare groups to test for item bias. As a result, future research on the JCQ may want to focus more on detecting items that may be biased (Mellenbergh, 1989).

Summary

There are both advantages and disadvantages to implementing IRT. However, being able to analyze the measurement error of each of the items as well as the person characteristics is a significant benefit to developing and validating a scale or questionnaire. Thus, IRT has an advantage over CTT methods for the current study of validating the JCQ.

Once additional validity evidence has been accumulated for the JCQ using IRT, some exciting new developments will be able to be made in the area of work stress. Researchers would be able to be more consistent when it came to test development, inspecting item bias, and the use of adaptive testing. In all, using IRT to accumulate validity evidence for the JCQ is promising not only in the present day, but in the application of the JCQ in future research. This is especially important since work stress and the health of the worker has become globally important (Briner, 2000).

CHAPTER III

Introduction

The following section introduces the methodology used in the current study including research design, research questions, sample, data collection, ethics, instruments, and data analysis. It concludes with an introduction into the procedures used for both the factor analysis as well as, the IRT analysis (i.e., using Winsteps).

Methodology

Research Design and Research Questions

The current research study is a validity study using quantitative methodologies. As the data was scored polytomously (e.g. strongly agree to strongly disagree) using a graded response model (i.e., sequentially scored data), the one-parameter IRT program Winsteps, was used to accumulate additional validity evidence for the JCQ. More specifically, the following research questions were examined:

1. What is the dimensionality of the JCQ?
    a. What factors and associated items constitute the JCQ?
    b. How well do the items fit each of the resulting subscales?
2. Utilizing Winsteps,
    a. How well does the data fit the model?
    b. How well do the items represent the latent trait (i.e., item quality)?
        i. How does where the range of items fall compare to where the person statistics fall on the latent trait?
        ii. When representing the latent trait, which items overlap and which items are overly spaced?
    c. What are the response probabilities for the polytomous items?
    d. What are the item difficulty placements on the latent trait?

However, in order to utilize IRT, the assumptions of the theory must first be met. The most important assumption of item response theory is the assumption of unidimensionality. Thus, EFA was employed to address the question:

1. What is the dimensionality of the JCQ?

Sample and Data Collection

Sample. The sample was taken from Janzen's (2006) research on "Gender, work, family and health". This sample is the most recent data accumulated on the JCQ from Saskatchewan. Janzen's (2006) study focused on the relationship between work, family and health status, and how gender (i.e., combination with life stage and economic circumstance) is a key issue in the determination of health. As a result Janzen (2006) proposed, based on recent labor force and social developments in Canada that multiple roles and health needed to be examined. Several scales were used to measure the level of well being, including the JCQ, in order to provide information for policy interventions (Janzen, 2006).

The sample consisted of Saskatchewan residents who were employed parents with at least one child under the age of 20 living at home. They also had to speak English and be employed full-time or part-time. Approximately 100 participants were randomly selected from each job stratum (i.e., low, medium, and high), each gender (i.e., male and female) and each age category (i.e., 25-34 and 35-50) to improve sample heterogeneity. The total number of participants was N=1160 with 486 (41.9 %) males and 674 (58.1%) females, with a mean age of 36.0 (SD = 0.21). A total of 578 participants were between the ages of 25-34 and 582 participants were between the ages of 35-50. More individuals were in a relationship (i.e., 67.2%, n=779; n= 641, married; and n=138 living with a partner), than those who were not with a partner (i.e., 32.8%, n=381, n=20 widowed; n=96 separated; n=42 divorced; n=123 single). In regards to occupation type; 200 (17.1%) were employed as managers or professionals, 171 (14.7%) were employed in teaching or in teaching-related roles, 178 (15.3%) were employed in the medical and health field, 450 (38.8%) were employed in clerical/sales/service, 52 (4.5%) were employed in construction trades, 39 (3.4%) were self-employed, 26 (2.2 %) were employed in the transportation field, 26 (2.2%) were civil servants, 2 (0.2%) were farmers, and 16 (1.4%) did not respond (Janzen, 2006; See Appendix D for a summary of the frequencies completed).

Data Collection. Selected participants were reached by trained telephone interviewers who randomly dialed the participant's number and first made sure that they fit the sample criteria (i.e., English speaking employed full or part-time, having at least one child at home). The interviews were approximately 40 minutes in length and data was collected by the Computer Assisted Telephone Interviewing (CATI) System. A trained local market company used the CATI system to contact participants when the time was convenient for them (Janzen, 2006).

Ethics. Although the current study is a secondary data analysis, ethics approval was obtained to conduct the additional analyses. The original data was collected by Janzen (2006) and no new participant data was required for the present study.

Instruments

Demographic Questions: Demographic questions were obtained from the National Population Health Survey (NPHS; Statistics Canada, 1996). Demographic variables included age, gender and educational attainment (Janzen, 2006).

Job Content Scale. Although Janzen (2006) examined many aspects of an employee's internal and external environment, only one scale, Karasek and Theorell's (1990) Job Content Questionnaire, version 1.11, was utilized in the current study. However, data was not collected for the Skill Utilization question (i.e., part of the Decision Latitude Scale), 2 items from the Coworker Social Support scale, the Physical Job Demands and 3 items from Job insecurity scale. Thus, there were a total of 6 items from the Skill Utilization scale, 3 items from the Decision Authority scale, 9 items from the Psychological Demands scale, 5 items from the Supervisor Social Support scale, 4 items from the Coworker Social Support scale (i.e., from a total of 6 items), and 3 items from the Job Insecurity scale (i.e., from a total of 6 questions). Only four scales were examined. Table 3 presents JCQ recommended version 1.11, which is polytomous in nature, with the abbreviated wordings for each of the scales. Although Janzen (2006) calculated reliabilities for each subscale using Cronbach's alpha, no other psychometric analyses were conducted. Furthermore, no new questions were added to the subscales, and no questions were modified. Missing items were coded as missing, and used with both the EFA and IRT analysis. Thus, the psychosocial workplace environment was measured using items from the recommended version (1.11) of the JCQ. The items used from the JCQ recommended version 1.11, can be found in Appendix E.

Polytomous items. The JCQ recommended version 1.11 contains 41 items, three scales (i.e., Decision Latitude, Psychological Demands, and Social Support), and four subscales (i.e., Decision latitude: Decision Authority and Skill Discretion; Social Support: Coworker Social Support and Supervisor Social Support. Each scale (and subscale) is measured by a 4 point response scale (i.e., strongly agree, agree, disagree, and strongly disagree; Karasek et al., 1998). As part of Janzen's (2006) study, a fifth response was added (i.e., unsure) and only four scales were analyzed.

The Decision Latitude (e.g. my job requires that I learn new things; 9 items) and Psychological Demands (e.g. my job requires working very hard; 9 items) scales try to predict the worst unfavorable effects of psychological strain when the worker's demand is high and Decision Latitude (i.e., independence to make decisions) is low. Skill Discretion (i.e., 6 items) and Decision Authority (i.e., 3 items) subscales make up the Decision Latitude scale and are highly correlated. Skill Discretion is measured by the amount of creativity and skill (i.e., variety; item: my job requires me to be creative) that a worker is required to employ on the job. On the other hand, Decision Authority is measured by how much a worker is able to make decisions on his or her own (i.e., autonomy; item: on my job, I have very little freedom to decide how I do my work; Karasek et al., 1998)

The Social Support scale includes two subscales; Coworker Social Support (i.e., 4 items) and Supervisor Social Support (i.e., 5 items). Jobs that are high in demands, low in support, and low in control, carry the highest risk of psychological strain. The Coworker Social Support scale (e.g. People I work with are competent in doing their jobs) ask different areas of how coworkers are supportive. On the other hand, the Supervisor Social Support scale (e.g. my supervisor pays attention to what I am saying) also tries to measure how supportive the supervisor is (Karasek, et al., 1998).

Finally, the Job Insecurity scale contains 3 items. Items like "My job security is good" try to measure the ability to adapt to a changing labor market. This scale can depend on the requirements of the labor market (e.g. particular skills), and possibility of future career development possibilities (Karasek et al., 1998).

Data Analysis

In order to accumulate additional validity evidence for the JCQ, a secondary data analysis was conducted using IRT. The data was analyzed using Winsteps, which is a one parameter IRT Rasch model. Rasch measurement is based on the Guttman scalogram, which also measures unidimensionality. However, prior to running a factor analysis and an IRT analysis, frequencies were run on the data and missing data was coded as missing. Following this, factor analyses were conducted to ensure that the assumptions of IRT were met.

There are two assumptions underlying the IRT model. These include; (1) unidimensionality and (2) local independence (Hambleton et al., 1991). These assumptions should be met in order to correctly fit data to a model. In instances where the assumptions are not

met; the model will be questionable as to whether the application of the model to the data is appropriate (Hambleton & Swaminathan, 1985). To make certain that local independence, but more specifically that unidimensionality is met, requires that the items are first able to fit the model (i.e., how well the interrelationships predicted by the model correspond to the ones actually observed); this is also called goodness of fit. Once completed, the second assumption is assumed to be met, since IRT is based on probability (Zenisky, Hambleton, & Sireci, 1999)

Factor Analysis

A confirmatory factor analysis (CFA) was not completed on the JCQ recommended version 1.11 as no other study, up until now, had previously examined this particular scale using these particular items and that there is inconsistency surrounding existing JCQ validity evidence. Thus, an exploratory factor analysis design was employed. To estimate a particular model using EFA, three sets of parameters must be identified; (1) factor loadings, (2) factor interrelationships, and (3) measurement errors (Norman & Streiner, 2003). The current study used SPSS for Windows (2006) to conduct the EFA.

Extraction. The first step was to perform an extraction to determine the number of latent factors inherent in the 30-item JCQ. The principle components extraction method, which is the most widely used extraction method (Gorsuch, 1983), was conducted followed by the Kaiser Guttman rule (i.e., retaining factors with eigenvalues that are greater or equal to 1.0), and Cattell's (1966) scree plot criteria were used to determine the number of factors. In addition, image factoring, followed by a varimax rotation, was employed. The principle of image factoring extraction is that it minimizes residual images, whereas both principle components and principle axis extraction techniques maximizes the amount of variance accounted for (Gorsuch, 1983).

Rotation. . Principle axis analysis was selected because it produces more conservative loadings than principal components analysis (Gorsuch, 1983), which is appropriate in the scale construction context when the purpose is to select the best fitting items. Factor rotations and transformations were performed and both orthogonal (i.e., unrelated; varimax) and oblique (i.e., related; direct oblimin) solutions were assessed.

Item Response Theory Analysis

In order to utilize the IRT based program (i.e.,Winsteps) with the data obtained from Janzen (2006), a number of steps were taken to review the items and scales comprising the JCQ.

These steps included examining (1) the dimensionality of the JCQ; (2) overall model fit to the data; (3) diagnosis statistics (i.e., item polarity, empirical category-item measures, dimensionality, and item misfit); and (4) the item probability curve, the item map and the ICC.

Overall Model fit. To examine overall model fit, the Winsteps summary statistics (i.e., diagnosis statistics) were examined. Specifically, fit statistics were examined in order to assess how closely the actual data lined up with what the measurement system predicts (i.e., produces fit statistics from response residuals). Two chi square ratios are computed and are labeled infit and outfit (i.e., both measured by mean square statistics; MNSQ's) statistics (i.e., chi square statistic divided by their degrees of freedom; t-statistic). Outfit statistics refer to the responses that are far off from what is expected (i.e., outliers) and are also less a threat to measurement. Infit statistics are influenced by response patterns, are usually hard to diagnose and are a greater threat to measurement. Bradley and Sampson (2005) suggest that infit and outfit measures below -2.0 and above +2.0 indicate less compatibility with the model. The chi-square outfit statistic is the recommended statistic to report (Linacre, 2006).

The MNSQ (i.e., mean square infit and outfit statistic) has an expected value of 1. 0. Linacre (2006) has several recommendations for interpreting the MNSQ values; (1) Values > 2.0 distorts the measurement system; (2) Values between 1.5-2.0 are unproductive for the construction of measurement; (3) Values between 0.5- 1.5 are productive for measurement; and (4) Values < 0.5 are less productive and may provide misleadingly good estimates of reliabilities and separation measures. High mean squares (i.e., MNSQ; positive t-squares) are a greater threat to validity than low mean squares because values higher than 1.0 (i.e., underfit) indicate that there is noise and another source of variance. On the other hand, low MNSQ's (i.e., values lower than 1.0) indicate overfit and that the model predicts the data too well. A value >1.5 suggests a divergence from unidimensionality of the data and not the measure (Linacre, 2006).

The Model fit table provided by Winsteps contains person and item statistics including the mean and standard deviations of the raw score, the count (i.e., the number of times a response was used by an individual), the measure (i.e., logit conversions of the raw score), and infit (i.e., information weighted fit statistics) and outfit statistics (i.e., outlier- sensitive fit statistic). Logit conversions are completed when comparisons are made within constructs containing different sets of item types (e.g. agree, strongly agree, etc). Logits are also the natural unit for the logistic ogive. The ZSTD (i.e., infit and outfit z standard deviation) is the t-standardized value to

demonstrate a theoretical model with a mean of 0.00 and standard deviation of 1.00 (Linacre, 2006).

The reliability and separation measures provide information about the consistency of the items or persons. Person reliability can depend on sample ability variance, length of test, number of categories per item, and sample-item targeting. On the other hand, item reliability can depend on item difficulty variance and person sample size (i.e., large sample = high item reliability). Person reliability is also regarded as the traditional "test" reliability, as low values are indicative of a narrow range of person measures (Linacre, 2006). Person reliability refers to how well the test discriminates the sample into enough levels for the researcher's purpose, whereas item reliability refers to how well the items are located on the latent variable.

Separation is defined as the spread of positions (i.e., either item or person), which is also considered the "test reliability" (Bradley & Sampson, 2005). Since test reliability is made up of 'true score" and "error score", the separation measures assess the "error score". The person separation identifies if the rating scale discriminates between persons, while the item separation identifies if the items are producing a well fitted variable (i.e., less error variance; Wright, 1996) The separation measures are the ratio of the adjusted person or the item, the "true" standard deviation, to the square-root of the average error variance (i.e., Root mean square/the persons or items), the error standard deviation (Linacre, 2006). In all, low separation reliability is better.

Diagnosis Statistics

The diagnosis statistics were assessed in order to investigate, via a step by step procedure, the results of the analysis. The diagnosis statistics used in this analysis include, item polarity, empirical item-category measures, dimensionality, and item misfit.

Item Polarity. Item polarity was examined to ensure that the items were in the same direction as the latent trait. Hence, all the items should have positive correlations. If not, negative values will indicate problems with the items. If values are negative this would indicate the rating scale items are in a reversed direction or items were mis-scored (Linacre, 2006).

Empirical Item-Category Measures. Each construct was reviewed in order to determine whether the response categories (i.e., strongly agree to strongly disagree) were aligned in the same direction (Linacre, 2006). A graphical display is provided and it is essential that the response categories (i.e., labeled from 1-5) are in place from 1 to 5 from left to right (Sampson & Bradley, 2005)

Dimensionality. Dimensionality was investigated to report any evidence of convergent validity. This was examined to ensure that the items are assigned to the same dimension. The analysis pinpoints secondary dimensions while completing a principle components/contrast of the observed residuals. Linacre (2006) indicated not to use the Rasch-residual-based Principal Components Analysis (PCAR; unrotated principle components) as the usual factor analysis, as the Rasch based model only examines variances and contrasts. PCAR hopes to accomplish the opposite of what is achieved by a factor analysis. That is, the least number of contrasts are found to explain the most variance as possible (Linacre, 2006).

To assess the variance, Winsteps produces a standardized residual variance scree plot. It displays the total variance in observations (T), variance explained by measures (M; variance explained by item difficulties, person abilities, and rating scale structures), and unexplained variance (U). Thus, one would think that the more the variance is explained by measures the more it is unidimensional. Yet, this is not the case. For a Rasch model, the more unidimensionality depends on the size of the second dimension in the data. It also provides unexplained variance by first, second, etc contrast. It is simply variance that is unexplained by the measures that can be explained by a contrast (Linacre, 2006). In all, the purpose of the PCAR does not construct variables, but rather explains variance.

Linacre (2006) gives some suggestions for examining variance; (1) > 60% is good for variance explained by measures; (2) <3.0 is good for unexplained variance explained by the 1st contrast; and finally (3) <5% is good for unexplained variance explained by 1st contrast. However, he also clarifies that there are many exceptions. As a final note, Linacre (2006) also states that the empirical explained variance should be close to the modeled explained variance.

Unidimensionality, which is a part of the steps in the IRT analysis, is also an assumption of IRT. Before the methods outlined above, an EFA was conducted to ensure that the items met the assumption of unidimensionality, within the five constructs.

Item Misfit. Means and mean squares were examined to determine if there were any items that were really large or really small (i.e., items that do not fit = not close to 1.0). Items that are not close to 1.0 indicate how much that item misfits the Rasch model .Thus, item misfit identifies misbehaving items. Misbehaving items are identified as having large mean squares (Linacre, 2006).

Probability Curve (Item)

Only the item probability curve is important for this analysis, as the current study is only interested in the item measures rather than the person measures. The probability curve displays the possibility of each response on a measurement continuum (Linacre, 2006). Tejada and Rojas (2005) state that there are two areas of interest for examining the probability curves (i.e., category probability curves); (1) It provides information as to the response alternatives; and (2) the intersections between the curves define the most probable responses. Each region will illustrate the highest probability of responding to that specific response category. As such, it also enables the researcher to predict responses (Bradley & Sampson, 2005)

Item Map

The item map displays respondents and items against each other and is a determinant of item quality. A researcher can utilize the item map to check for gaps (i.e., indicating more items could be developed to better represent the variable), and where respondents are in comparison to the items (i.e., for polytomous items: if items are above the respondents, this would indicate difficult items to indorse; items below respondents, indicate easier items to indorse). If the items do not match up with the intended respondents, then additional items may need to be developed so there would be a better picture of the sample on the construct (Sampson & Bradley, 2005). Therefore, items above the respondents indicate the items are difficult to indorse for the matched sample, which indicate that other items need to be constructed so that they can be arranged lower on the map (Bradley & Sampson, 2005).

Item Characteristic Curve

As mentioned previously, the ICC displays the relative frequency of each measured level and each item. They are measured in logits and displayed in a figure. The vertical axis displays the probability of a correct response, whereas the horizontal axis presented the ability level of the respondent (Linacre, 2006). The ICC displays the probability of answering a certain item on each of the latent traits of the five constructs. The forms of the ICC will change as the amount of parameters chosen changes (Ellis, Becker, & Kimmel, 1993). The parameters are defined as a_i (i.e., item discrimination), b_i (i.e., item difficulty) and c_i (i.e., item guessing). However, since a one parameter model is being used only the a_i parameter will be observed (Botempo, 1993).

Summary

The current study utilized a secondary data analysis to conduct a factor analysis and a Winsteps analysis on the JCQ. A total of 1160 individual's participated and 4 constructs were examined (i.e., Decision Latitude, Psychological Demands, Social Support, and Job Insecurity) from the JCQ recommended version 1.11. The following chapter provides an examination of the results with a discussion of these results provided in the last chapter of the thesis.

Chapter IV: Results

Introduction

The following section presents the results for both the factor analysis and the Winsteps analysis (i.e., IRT analysis). As mentioned previously, an EFA was conducted. Following the extraction of an optimal number of factors to satisfy the assumption of dimensionality, the resulting factors were rotated. A number of steps are outlined to describe how the JCQ was analyzed via Winsteps.

Factor Analysis Results

In order to ensure that the assumption of unidimensionality was met, an EFA was conducted. Although the JCQ is theorized to consist of five scales (Karasek et al., 1998), previous research had utilized different sub-scales, and identified different factor patterns depending on which items were included. Furthermore, because no previous research has applied IRT to the JCQ, the current study implemented an EFA methodology to determine the unidimensionality (i.e., prior to conducting IRT).

Extraction and Rotation

Extraction. The principle components extraction, of the 30 items, revealed that there were seven possible factors using the Kaiser-Guttman rule (Kaiser, 1960). In contrast, Cattell's scree plot criteria (Cattell, 1966) indicated two to three possible factors. The image factoring extraction identified 7 possible factors, with only 6 containing eigenvalues above 0.30[ii] (see Appendix G).

Rotation. An orthogonal solution was achieved through principle axis followed by a varimax rotation. An oblique solution was accomplished by completing a principle axis extraction followed by a direct oblimin rotation.

Initial solution

Following the rules of simple structure and Gorsuch's (1983) 0.30 criteria for minimum loading pattern interpretations, an oblique rotation (i.e., using a principle axis followed by a direct oblimin rotation) was found to best represent the 30 items. Fewer doublets (i.e., items that were significant on more than one factor) were identified when oblique rotations were performed. However, due to an underlying hierarchical structure (i.e., further removal of items and exploratory factor analyses revealed that items were still moving), multiple rotations and extractions (i.e., using the same process described above) were completed on separate scales that

were revealed in the analysis for the 30 items (i.e., Decision Latitude, Psychological Demands, Supervisor Social Support, Coworker Social Support, and Job Insecurity). The principle axis extraction followed by direct oblimin rotation provided the best results for each of the Skill Discretion and Decision Authority items, the Psychological Demands items, the Social Support items, and the Job Insecurity items.

Skill Discretion and Decision Authority

Table 4 demonstrates both the Skill Discretion and Decision Authority items load best on one factor, with the exception of 3 items. The overall results revealed that Skill Discretion item 1 loaded on factor 2 and 3, Skill Discretion item 2 loaded on factor 2, and Decision Authority item 2 loaded on factor 3. Once removed, unidimensionality for the Skill Discretion and Decision Authority scale were met. Items 3-6 of the Skill Discretion scale and items 1 and 3 of the Decision Authority scale are included in the final scale. These items (Skill Discretion: (3) "My job requires me to be creative"; Skill Discretion: (4) "My job requires a high level of skill"; Skill Discretion: (5) "I get to do a variety of different things on my job"; Skill Discretion; (6) "I have an/the opportunity to develop my own special abilities"; Decision Authority: (1) "My job allows to make a lot of decisions on my own"; Decision Authority: (3) "I have a lot of say about what happens on my job") appear to be better representing the construct and so was renamed Decision Latitude.

Psychological Demands

Principle axis extraction followed by a direct oblimin rotation was conducted on the 9 Psychological Demands items. Extraction results (see Table 5) indicated a 3 factor solution with the Psychological Demands item 3 and 4 loading on the second factor, the Psychological Demands items 1 and 2 loading on the third factor, and the remaining items loading on the first factor. Following the removal of the first four items, and re-factor analyzing the remaining 7 items to check for unidimensionality, unidimensionality was achieved. Five Psychological Demands items remained (Psychological Demands: (5) "The demands that other people make of me often conflict"; Psychological Demands: (6) "My job requires long periods of intense concentration on the task"; Psychological Demands: (7) "My tasks are often interrupted before I can finish them so that I have to back to them later"; Psychological Demands: (8) "My job is very hectic"; Psychological Demands: (9) "Waiting on work from other people or departments often slows me down on my job") with factor loadings above 0.496.

Table 4

Principle axis extraction results for the decision latitude items, with a direct oblimin transformation.

	Factor		
	1	2	3
Skill Discretion: (1) My job requires that I learn new things		-.359	.513
Skill Discretion: (2) My job involves a lot of repetitive work			.477
Skill Discretion: (3) My job requires me to be creative	.729		
Skill Discretion: (4) My job requires a high level of skill	.729		
Skill Discretion: (5) I get to do a variety of different things on my job	.761		
Skill Discretion: (6) I have an/the opportunity to develop my own special abilities	.872		
Decision Authority: (1) My job allows me to make a lot of decisions on my own	.816		
Decision Authority: (2) On my job, I have very little freedom to decide how I do my work		.719	
Decision Authority: (3) I have a lot of say about what happens on my job	.702		

Table 5

Principle axis extraction results for the psychological demand items (i.e., items 5-9) with a direct oblimin extraction.

	Factor		
	1	2	3
Psychological Job Demands: (1) My job requires working very fast			-.762
Psychological Job Demands: (2) My job requires working very hard			-.774
Psychological Job Demands: (3) I am not asked to do too much work		.567	
Psychological Job Demands: (4) I have enough time to get the job done		.764	
Psychological Job Demands: (5) The demands that other people make of me often conflict	.513		
Psychological Job Demands: (6) My job requires long periods of intense concentration on the task	.496		
Psychological Job Demands: (7) My tasks are often interrupted before I can finish them so that I have to go back to them later	.684		
Psychological Job Demands: (8) My job is very hectic	.676		
Psychological Job Demands: (9) Waiting on work from other people or departments often slows me down on my job	.624		

Social Support

Principle axis extraction of the social support items (i.e., Coworker Social Support items 1-4 and Supervisor Social Support items 1-5) resulted in a 3 factor solution (see table 6). One item (i.e., Supervisor Social Support item 5) loaded on the third factor and was removed. However, Supervisor Social Support item 4 loaded on both the first and third factor. A re-factor analysis of the 4 Social Support items resulted in one factor being extracted. Factor 1 consisted of 4 Supervisor Social Support items (i.e., Supervisor Social Support: (1) "My supervisor is concerned about the welfare of those under him/her"; Supervisor Social Support: (2) "My supervisor pays attention to what I am saying"; Supervisor Social Support: (3) "My supervisor is helpful in getting the job done"; Supervisor Social Support: (4) "My supervisor is good at getting people to work together") with factor loadings ranging from 0.386 to 0.898. One Supervisor Social Support item (i.e., Supervisor Social Support: (5) "I am exposed to hostility or conflict from my supervisor") did not load on the first factor and was removed.

The Coworker Social Support scale contained four items (i.e., Coworker Social Support: (1) "People I work with are competent in doing their jobs; Coworker Social Support": (2) "People I work with take a personal interest in me"; Coworker Social Support: (3) "People I work with are friendly"; and Coworker Social Support: (4) "People I work with are helpful in getting the job done").

Job Insecurity

A separate EFA was conducted on the Job Insecurity scale. No items needed to be removed as the first extraction and rotation met the assumption of unidimensionality. Three items made up the Job Insecurity construct (i.e., Job Insecurity: (1) "My job security is good"; Job Insecurity: (2) "My prospects for career development and promotions are good"; and Job Insecurity: (3) "In five years, my skills will still be valuable").

Once the scales appeared to be unidimensional, internal consistencies (i.e., Cronbach's alpha) were calculated for each of the subscales. Table 7 provides the scales, the items included, and the reliabilities of each scale. Four of the scales had reliability estimates over 0.70, which is considered a good estimate of internal consistency (Nunnally & Bernstein, 1994). Although, the Job insecurity scale had a reliability estimate below 0.70, this is likely due to the fact that the scale only contained 3 items instead of the 6 items from the complete scale.

Table 6

Principle axis extraction results of the social support items, with a direct oblimin extraction.

	Factor		
	1	2	3
Co-worker Social Support: (1) People I work with are competent in doing their jobs		.470	
Co-worker Social Support: (2) People I work with take a personal interest in me		.696	
Co-worker Social Support: (3) People I work with are friendly		.880	
Co-worker Social Support: (4) People I work with are helpful in getting the job done		.692	
Supervisor Social Support: (1) My supervisor is concerned about the welfare of those under him/her.	.720		
Supervisor Social Support: (2) My supervisor pays attention to what I am saying.	.898		
Supervisor Social Support: (3) My supervisor is helpful in getting the job done.	.819		
Supervisor Social Support: (4) My supervisor is good at getting people to work together.	.386		-.447
Supervisor Social Support: (5) I am exposed to hostility or conflict from my supervisor.			.612

Table 7

Internal consistencies, using Cronbach's alpha, of the items for the Winsteps analysis.

Scale	Items	# of items	α	Valid N	N
Decision Latitude	3-6; 1,3	6	0.897	1160	1160
Psychological Demands	5-9	5	0.759	1160	1160
Coworker Social Support	1-4	4	0.769	1115	1160
Job Insecurity	1-3	3	0.648	1160	1160
Supervisor Social Support	1-4	4	0.832	686	1160

Winsteps Results

Once dimensionality of the items was established using factor analysis, the Winsteps program was used to examine each of the five constructs; (1) Decision Latitude (i.e., Skill Discretion, 3-6; Decision Authority, 1, 3); (2) Psychological Demands items (i.e., 5-9); (3) Coworker Social Support items (i.e., Coworker Social Support items 1-4); Supervisor Social Support items (i.e., 1-4); and (5) Job Insecurity (i.e., 1-3). The following section reviews the IRT results in relation to the five constructs and provides overall model fit, diagnosis statistics (i.e., item polarity, empirical item-category measures, dimensionality, and item misfit), item probability curve, item map, and the ICC.

Overall Model Fit

Overall model fit for the constructs appear in table 8. They display separation reliability, item reliability, infit and outfit statistics and sample size. Results demonstrate that both infit and outfit MNSQ measures for both the person and item statistics are close to what is expected (i.e., 1.0; Linacre, 2006) for each construct. The fifth construct (i.e., Job Insecurity) received the lowest MNSQ value. However, both MNSQ (i.e., outfit and infit) are below 1, indicating that there may be some dependency between items (i.e., lacking independence). Yet, as Linacre (2006) advised items with MNSQ's between 0.5-1.5 are productive for measurement and as such, each constructs' MNSQ's were between those criteria.

Both person reliability (i.e., 0.78; 0.71; 0.63; 0.70; 0.53) and item reliability (i.e., 0.97; 0.98; 0.96; 0.97; 1) are satisfactory and good estimates (Streiner, 2003), except for the fifth (i.e., Job Insecurity) construct which was lower than 0.6. In this case, low item reliability (or person reliability) indicate that there are a narrow range of item measures or it was too small a sample. In this case, both low reliabilities were related to the person statistics, thus, a larger sample may be needed to test these items (Linacre, 2006).

Separation is the ratio of true variance to observed variance (Linacre, 2006). Thus, a larger number indicate a larger difference in variance. Person separation estimates for the five constructs (i.e., 1.86; 1.58; 1.3; 1.52 and 1.07) were quite small indicating a small difference (i.e., IRT analysis identifies on average 1 performance strata or ability). On the other hand, item separation for the five constructs (i.e., 5.71; 6.83; 5.21; 5.92; 15.64) indicated that some had a larger variance, especially for the fifth construct (i.e., Job Insecurity), which was 3 times larger than the rest. In all, the IRT analysis identified 5 performance strata (or error variance) for

Table 8

Overall model fit for constructs #1 - #5 (i.e., decision latitude; psychological demands; coworker social support; supervisor social support; job insecurity).

Construct 1		N	Infit		Outfit	
	Persons	1160	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		0.99	-0.1	0.98	-0.1
	S.D.		0.75	1.3	0.75	1.3
	Separation; Reliability	1.86; 0.78				
	Items	6	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		0.99	-0.3	0.98	-0.5
	S.D.		0.15	3.0	0.17	3.6
	Separation; Reliability	5.71; 0.97				
Construct 2		N	Infit		Outfit	
	Persons	1160	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		0.98	-0.2	0.97	-0.2
	S.D.		0.95	1.4	0.94	1.4
	Separation; Reliability	1.58; 0.71				
	Items	5	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		0.99	-0.2	0.97	-0.8
	S.D		0.11	2.7	0.12	2.9
	Separation; Reliability	6.83; 0.98				

Table 8 continued

Overall model fit for constructs #1 - #5 (i.e., decision latitude; psychological demands; coworker social support; supervisor social support; job insecurity).

Construct 3		N	Infit		Outfit	
	Persons	1160	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		0.91	-0.5	0.91	-0.5
	S.D.		1.59	1.5	1.6	1.5
	Separation; Reliability	1.3; 0.63				
	Items	4	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		1	-0.3	0.96	-1.1
	S.D.		0.22	4.5	0.2	3.9
	Separation; Reliability	5.21; 0.96				
Construct 4		N	Infit		Outfit	
	Persons	1160	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		0.92	-0.4	0.94	-0.4
	S.D.		1.23	1.4	1.3	1.5
	Separation; Reliability	1.52; 0.7				
	Items	4	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		0.99	-0.4	0.94	-1.0
	S.D		0.28	4.6	0.29	3.3
	Separation; Reliability	5.92; 0.97				

Table 8 continued

Overall model fit for constructs #1 - #5 (i.e., decision latitude; psychological demands; coworker social support; supervisor social support; job insecurity).

Construct 5		N	Infit		Outfit	
	Persons	1160	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		0.83	-0.3	0.82	-0.3
	S.D.		1.08	1.2	1.07	1.2
	Separation; Reliability	1.07; 0.53				
	Items	3	IMNSQ	ZSTD	OMNSQ	ZSTD
	Mean		1.01	0	0.85	-2.3
	S.D.		0.11	2.5	0.08	1.4
	Separation; Reliability	15.64; 1.00				

construct #1, #3, and #4, and 6 for construct #2. Consequently, construct #5 had 15 identifiable performance strata (i.e., or error variance).

Diagnosis Statistics

Item Polarity. As indicated by the point biseral correlations, all five constructs have positive correlations (see Appendix H for an example for the Supervisor Social Support scale; item polarity for the other constructs are available from the author). Construct #1 (i.e., Decision Latitude) had the following correlations; Decision Authority (3): 0.74; Skill Discretion (4): 0.79; Skill Discretion (3): 0.79; Skill Discretion (5): 0.80; Decision Authority (1): 0.81; and Decision Authority (3): 0.83. Construct # 2 had the following correlations; Psychological Demands (5): 0.68; Psychological Demands (6): 0.69; Psychological Demands (7): 0.74; Psychological Demands (8): 0.73; and finally Psychological Demands (9): 0.68. Construct #3 had the following correlations; Coworker Social Support (1): 0.64; Coworker Social Support (4): 0.72; Coworker Social Support (3): 0.73; and Coworker Social Support (2): 0.75. Construct #4 had the following correlations; Supervisor Social Support (4): 0.72; Supervisor Social Support (1): 0.78; Supervisor Social Support (3): 0.79; and Supervisor Social Support (2); 0.81. Finally, construct # 5 (i.e., Job Insecurity) had the following correlations; Job Insecurity (3): 0.67; Job Insecurity (1): 0.69; and Job Insecurity (2): 0.80. This signifies that all the items in each construct are in the same direction as the latent trait.

Empirical Item-Category Measures. Empirical item-category measures for the five constructs were examined (see Appendix G for an example for the Supervisor Social Support scale; empirical item-category measures for the other constructs are available from the author). For the first (i.e., Decision Latitude) all response items were ordered from right to left (i.e., 1-4). In comparison, for the other four constructs (i.e., Psychological Demands, Coworker Social Support, Supervisor Social Support, and Job Insecurity); there were disordered items (i.e., 1 3 2; "less frequently observed intermediate categories"; Linacre, 2006).

All support constructs, as well as, the Psychological Demands construct, had disordered items. Two of the Psychological Demand items (i.e., 5 and 8) had disordered items (i.e., 1 2 3 5 4). Four of the Coworker Social Support items (i.e., 2, 4, and 1) had disordered responses (i.e., 1 2 3 5 4), whereas Coworker Social Support item number 3 did not (i.e., 1 2 3 4 5). Supervisor Social Support item number 1 (i.e., 1-4) and 4 (i.e., 1-4) for the fourth construct both had ordered responses. Conversely, Supervisor Social Support item number 3 (i.e., 1 5 2 3 4) and number 2

(i.e., 1 2 3 5 4) had disordered responses. Lastly, all the Job Insecurity items (i.e., fifth construct) had disordered responses (i.e., 1 2 3 5 4; 1 2 3 5 4; 1 2 5 3 4).

Dimensionality. Dimensionality figures were examined (see Appendix H for the table of each construct). For the first construct (i.e., Decision Latitude) variance explained by the measures was 68.3%, while the variance unexplained was at 31.7%. For construct #2 (Psychological Demands) the variance explained by measures was 64.0 %, while the unexplained variance was 36.0%. Construct #3 (Coworker Social Support) was low with the variance explained by the measures at 55.7% and the unexplained variance at 44.3 %. For construct #4 (Supervisor Social Support) and construct #5 (Job Insecurity) the variance explained by the measures was high (i.e., 85.8%; 85.7%), while the unexplained variance was low (i.e., 14.2%; 14.3%). In all, the explained variance was high for construct #4 and #5, while the unexplained variance was high for construct #1, #2 and #3.

After comparing the empirical explained variance to the modeled explained variance it was noted that each construct had empirical values which were close to the modeled values (i.e., explained variance if the model fit the data). Also, each construct had empirical values < 3.0 for the unexplained variance in the 1st contrast, indicative from one of Linacre's (2006) guidelines.

Item Misfit. Item misfit tables for the five constructs were examined for any large MNSQ's (see Appendix I for an example for the Supervisor Social Support scale; item misfit tables for the other constructs are available from the author; Linacre, 2006). The first construct (i.e., Skill Discretion and Decision Authority) had two items with MNSQ's larger than 1.0 (i.e., Decision Authority item 3 and Skill Discretion item 3). The second (i.e., Psychological Demands) and fifth (i.e., Job Insecurity) constructs also had two items (i.e., Psychological Demand items 6 and 9; Job Insecurity item 1 and 3) with MNSQ's larger than 1.0. The third (i.e., Coworker Social Support) and fourth (i.e., Supervisor Social Support) constructs both had one item with a large MNSQ (i.e., Coworker Social Support item 1; Supervisor Social Support item 4).

Probability Curve

Category probability curves were examined for each item within each construct (see Appendix J for an example of the response probabilities from the Supervisor Social Support scale; probability curves for the other constructs are available from the author). Each item had five response categories (i.e., strongly disagree: 1, disagree: 2, agree: 3, strongly agree: 4, and

unsure: 5). the response probabilities for Construct #1 (i.e., Decision Latitude) were all the same. The probabilities were as follows; strongly disagree (most probable response seen at lower ability measures), strongly agree (most probable response seen at higher ability measures), and finally agree and disagree (middle ability level). Each item in construct #2 (i.e., Psychological Demands) had the following response probabilities; strongly disagree (most probable response seen at lower ability measures; easier), unsure (most probable response seen at higher ability measures; harder), strongly agree (at the higher end of the latent trait), and agree and disagree were roughly the same (middle ability level). Each item in construct #3 (i.e., Coworker Social Support) had the following response probabilities; strongly disagree (most probable response at the lower ability level), unsure (most probable response at the higher ability level), strongly agree (most probable response at the middle ability level), agree, and disagree (which was a lot lower than the rest). Each item in construct #4 (i.e., Supervisor Social Support) had the following response probabilities; strongly disagree (most probable response at the lower ability level), unsure (most probable response at the higher end of the ability), strongly agree (most probable response at the middle ability level), agree, and disagree. Finally, each item in construct #5 (i.e., Job Insecurity) had the following response probabilities; strongly disagree (most probable response at the lower end of ability), unsure (most probable response for the higher end of the ability scale), strongly agree (most probable response at the middle ability level), disagree, and agree. Overall, strongly agree was the most probable response in the middle of the ability level.

Item Map

The item maps for four of the constructs (i.e. construct #1- #5) are displayed in Appendix K. Explanations as to the visual representations of the maps are in the following paragraphs.

Construct #1: Decision Latitude. The item map for the first construct revealed that there were gaps in between the items, as well as, some of the items (i.e., Skill Discretion 3 and Skill Discretion 6; Decision Authority 1 and Skill Discretion 5) were representing the same part of the construct. Gaps ranged from 60-90 and 20-45 for the first construct (Decision Latitude). The items were also at the bottom of the where the people were responding, indicating easier items to indorse.

Construct #2: Psychological Demands. The item map for the second construct was a little more promising. Although the items were covering more of the latent trait, two items (Psychological Demands item #5 and Psychological Demands item #9) were covering the same

area of the latent trait. Each item was in the area (i.e., at the top) of where the respondents were situated (i.e., bulk of respondents ranged from 15-60). The gaps ranged from 60- 80 and (-) 10-40. These items were indicative of representing a better portion of the latent trait than the first construct.

Construct #3: Coworker Social Support. The item map for the third construct (i.e., Coworker Social Support) demonstrated items that also represented a good portion of the latent trait. However, in both the second and third construct other items would need to be developed to characterize the mass of respondents. No gaps were present between the items of the third construct, yet gaps ranged above and below the items (i.e., from 60-110 and 10-40). However, respondents ranged roughly from 20-80, while the items only ranged from 45-55.

Construct # 4: Supervisor Social Support and Construct #5: Job Insecurity. The item map for both the fourth and fifth construct had items that did not represent the latent trait very well. The fourth construct had items that were over the mass of respondents (i.e., indicating items too hard to indorse). Respondents ranged from (-)10 - 60 and there were missing areas from 60-90 and (-)10-40 for the items. There were also items represented the same part of the latent trait (i.e., representing the same part of the line; Supervisor Social Support item 2 and Supervisor Social Support item 1). Gaps were also present between the items (i.e., 50-60). The fifth construct also had gaps in between the items (i.e., Gap ranged between 50-55, for Job Insecurity items 2 and 3), indicating that more items need to be developed to represent the latent trait. Respondents ranged roughly from 8 – 85, whereas missing item gaps were present from 60-90, and (-) 10- 40.

Test Information: ICC

Each figure displays the increasing level of the underlying latent trait and performance on an item (McCarty, 2005). Since only a one-parameter model was used, only the b_i (i.e., item difficultly) can be observed for each item. The test taker has a 0.50 probability of responding correctly to the item on the rating scale. In contrast to dichotomous items where there is a right and wrong answer, polytomous items are based on difficult to indorse and easy to indorse (Linacre, 2006). Meaning, hard items may be confusing to the test taker or it is harder to agree to that item.

For construct #1 (i.e., Decision Latitude) the easiest item to indorse was Skill Discretion (4) “My job requires a high level of skill” and the most difficult item to indorse was Decision

Authority (3) “I have a lot of say about what happens on my job”. For construct #2 (i.e., Psychological Demands) the easiest item to indorse was Psychological Demands (8) “My job is very hectic” and the most difficult item to indorse was Psychological Demands (5) “The demands that other people make of me often conflict”. For construct #3 (i.e., Coworker Social Support) Coworker Social Support item 3 (i.e., “People I work with are friendly”) was easiest to indorse, while Coworker Social Support item 2 (i.e., “People I work with take a personal interest in me”) was the most difficult to indorse. The easiest item to indorse for construct #4 (i.e., Supervisor Social Support) was difficult to read as there were two items representing the same area of the construct (i.e., Supervisor Social Support (1): “My supervisor is concerned about the welfare of those under him/her”; and Supervisor Social Support (2): “My supervisor pays attention to what I am saying”). The most difficult item to indorse was Supervisor Social Support (4) “My supervisor is good at getting people to work together”. For construct #5 (i.e., Job Insecurity), Job Insecurity item # 3 (i.e., “In five years, my skills will still be valuable”) was easiest to indorse, while Job Insecurity item # 2 (i.e., “My prospects for career development are promotions are good” was the most difficult to indorse (see appendix L for an example).

Summary

Chapter IV presented the results of both the EFA and the IRT analysis. Numerous items (i.e., 8 items) had to be removed to ensure unidimensionality via factor analysis before IRT analysis could be performed. Results indicated that many of the items were representing the same area of the latent trait. The next section provides the discussion to the results.

CHAPTER V: Discussion

Introduction

This final section of the thesis will briefly review and summarize the main results found in chapter 4. The advantages and the disadvantage of using an IRT based method for examining the psychometric properties of a scale will be discussed, as well as, and limitations of the current study. This section will conclude with opportunities for future research.

Summary of Results

Using the items in the recommended version of the JCQ and an EFA to test the assumption of unidimensionality; a total of 5 constructs emerged from the 4 scales taken from the JCQ recommended version 1.11. The current study examined the recommended version (1.11) of JCQ with the following scales; (1) Decision Latitude: Decision Authority and Skill Discretion, (2) Psychological demands, and (3) Social Support. However, to conduct an IRT analysis, the assumption of unidimensionality must be met. One method is through factor analysis. However, results showed that the scales (and subscales) were not unidimensional, but after multiple extractions and rotations, 5 constructs emerged (i.e. Decision Latitude, Psychological Demands, Coworker Social Support, Supervisor Social Support, and Job Insecurity). The following section reviews the Winsteps results for the 5 constructs including the items that they were comprised of.

Construct # 1: Decision Latitude

The first construct encompassed Skill Discretion and Decision Authority and contained items “My job requires me to be creative”, “my job requires a high level of skill”, “I get to do a variety of different things on my job”, “I have an/the opportunity to develop my own special abilities”, “My job allows me to make a lot of decisions on my own”, and “I have a lot of say about what happens on my job”. Although 3 items were deleted because they violated the assumption of unidimensionality, the items that remained still had a very good internal consistency (i.e., $\alpha = 0.897$) indicating that that the items were reliable (Linacre, 2006). Conversely, the separation measures were large (i.e., person separation) which may suggest a larger error variance (i.e., 5 levels of error variance; Wright, 1996).

Overall model fit for the first construct (i.e., Decision Latitude) was low, in terms of MNSQ values (i.e., for both infit and outfit statistics), which signifies dependency among the items (i.e., no independence). However, the items were not substantially below 1.0, thus, it was

not a cause for concern (Linacre, 2006). Encouraging results were documented from the items including: (1) The polarities of the items were positive, which indicate they are in the same direction of the latent trait; and (2) Response items were ordered from left to right (i.e., with the highest response probability being strongly disagree and the lowest response probability being disagree). Conversely, when examining the item misfit, the construct had two items that had larger MNSQ's pointing toward how they do not fit within the construct.

As for the item map, items were largely representing the same area of the construct and did not contain enough spread (i.e., too low) to cover the person measures. The variance explained by the measures was good, but could have been higher, as it takes into account the measures based on item difficulties, person abilities and rating scale structures. On the other hand, the unexplained variance was high. Meaning, there was a high percentage of variance unaccounted for by the IRT analysis, but there was no effect of misfit or multidimensionality (i.e., the empirical explained variance is not extremely less than the modeled explained variance). The highest probable response at the lower ability levels was strongly agree, and the highest probable response at the higher ability level was strongly disagree. Overall, the diagnosis statistics suggested that the construct required more items, as well as, modification to the items that were already present to fully encompass the latent trait. Some items (i.e., Skill Discretion item # 3 and Skill Discretion item # 6; Decision Authority item # 1 and Skill Discretion item # 5), represented the same area of the latent trait (Linacre, 2006).

The ICC for construct #1 indicated that Skill Discretion item #4 was the easiest to indorse, while Decision Authority item #3 was the most difficult to indorse. This is also indicative with the item map, as more items are required to fully encompass the latent trait.

Construct #2: Psychological Demands

The second construct encompassed Psychological Demands. However, not all the items could be included (i.e., 1-4) as they violated the assumption of unidimensionality. The final construct contained the items "The demands that other people make of me conflict", "My job requires long periods of intense concentration on the task", "My tasks are often interrupted before I can finish them so I have to go back to them later", "My job is very hectic", and "Waiting on work from other people or departments often slows me down on my job". The item internal consistency was acceptable (i.e., Nunnally & Bernstein, 1994; $\alpha = 0.759$). However, the person reliability was low (i.e., the construct items does discriminate the sample into enough

levels; Linacre, 2006) and the item separation measures were high (i.e., above 1; indicated 6 levels of error variance; Wright, 1996).

Overall model fit pointed out MNSQ's below the accepted value. For low MNSQ's, it is suggested that there may be dependency among items, whereas for high MNSQ's it is suggested that there may be outliers. However, since the MNSQ's of the items were not substantially lower or higher than 1.0, there was no cause for concern (Linacre, 2006). Item measures for the second construct (i.e., Psychological Demands) were good; (1) item polarities were positive signifying the items are in the same direction as the latent trait; and (2) 4 response items were ordered from right to left, while only 2 items had disordered responses. The highest probable response at the lower ability level was strongly disagree, while the highest probable response at the higher ability level was unsure.

The variance explained by the measures was not very high, while the unexplained variance was high. This suggests that the items are not unidimensional in nature (Linacre, 2006). However, the second construct had two large MNSQ's, which signifies noise within the items (i.e., MNSQ's above 1.0 indicating outliers; Linacre, 2006). Item spread was better for the second construct, as the items were above the person measures. However, two items were representing the same area of the latent trait. This was also displayed in the ICC, as the easiest item to indorse was Psychological Demands #8 and the most difficult item to indorse was Psychological Demands item #5.

Construct # 3: Coworker Social Support

The third construct included four Coworker Social Support and no items needed to be deleted as the items met the assumption of unidimensionality. The items included "People I work with are competent in doing their jobs", "People I work with take a personal interest in me", "People I work with are friendly", "and People I work with are helpful in getting the job done". The internal consistency measure was satisfactory for the person reliability and good for the item internal consistency measure (i.e., 0.769). Person separation was low (i.e., 1 level of error variance), whereas item separation was higher (i.e., 5 levels of error variance).

Similar to the previous 2 constructs, overall model fit revealed low MNSQ's. As mentioned before, there may be dependency among the items (i.e., lack of independence). However, the MNSQ values were not substantially lower than 1.0, and thus, not a cause for concern (Linacre, 2006). The items were in the same direction as the latent trait, and compared to

the other constructs, only one item may be producing too much noise (i.e., possible outliers). Nevertheless, the items were disordered and only had a satisfactory percentage of explained variance by the measures (i.e., items do not point towards being unidimensional). The empirical item-category measures revealed disordered items for the third construct. Hence, the highest probable response at lower ability levels was strongly disagree, and the highest probable response at the higher ability level was unsure.

As for item spread, the items were relatively in the middle of the person measures and none were representing the same area of the construct meaning that they represent a better area of the latent trait. However, from the item map, more items could be developed to cover the lower end of the latent trait. The ICC identified the easiest item to indorse (i.e., Coworker Social Support #3) and the most difficult item to indorse (i.e., Coworker Social Support #2), which also displayed the spread of items across the latent trait.

Construct # 4: Supervisor Social Support

The fourth construct included the Supervisor Social Support items; "My supervisor is concerned about the welfare of those under him/her", "My supervisor pays attention to what I am saying", "My supervisor is helpful in getting the job done", and "My supervisor is good at getting people to work together". Only one item had to be removed (i.e., item #5) because it violated the assumption of unidimensionality. Internal consistency measures for the items were very good (i.e., $\alpha = 0.832$), while only satisfactory for the person measures. Thus, while the items are reliable, the person measures are slightly less reliable. The person separation value was low (i.e., 1 level of error variance), in comparison to the item separation, which was higher (i.e., 5 levels of error variance).

Overall model fit for the fourth construct revealed low MNSQ's, as with the previous constructs, which is explained as the items being too predictable (i.e., lacking independence). However, the values were not a cause for concern (i.e., not substantially lower than 1.0; Linacre, 2006). The item polarities were in the same direction as the latent trait and had very high explained variance (i.e., points towards unidimensionality; Linacre, 2006). On the other hand, the items were disordered and the item spread was satisfactory with items measuring the same area of the latent trait and only representing the top of the latent trait. The highest probable response at the lower ability level was strongly disagree, and the highest probable response at the higher ability level was unsure. The most difficult item to indorse was Supervisor Social Support item

#4 and the easiest item to indorse was Supervisor Social support item #2 and #1. These results suggest that more items need to be created, as well as, some of the items already developed correspond to the same area of the latent trait (i.e., social support item #1 and social support item #2).

Construct # 5: Job Insecurity

The fifth construct was comprised of three items from the Job Insecurity scale. They were; "My job security is good", "My prospects for career development and promotions are good", and "In five years my skills will still be valuable". Item reliability was perfect, making one question the items, whereas the person reliability was satisfactory. The reason for this could be too small a sample or a narrow range of item measures. Internal consistency was satisfactory (i.e., $\alpha = 0.648$), which could be because the construct only contained 3 items. Person separation was good (i.e., Low = 1 level of error variance); conversely, item separation was very high (i.e., large difference in variance; 15 levels of error variance, Linacre, 2006). The large amount of error variance should be a cause for concern for the three items that make up the Job Insecurity scale.

Overall model fit revealed that the lowest MNSQ values were for the fifth construct, in comparison to the rest. This would indicate that the items are very predictable (i.e., lacking independence between the items). However, the items were in the same direction as the latent trait and had a high percentage of explained variance (i.e., points towards unidimensionality). On the other hand, the items were disordered, the probable response categories were the same as constructs 2, 3 and 4 (i.e., highest probable response at the lower ability level: strongly agree; highest probable response at the higher ability level: unsure) and the spread of items to the person measures were incongruent. The items were being responded to at the high end of the latent trait, with a large spread between the most difficult item to indorse (i.e., Job Insecurity #2) and the easiest item to indorse (i.e., Job Insecurity #3).

Summary

In all, item spread was smaller than the person spread, as each construct was lacking items to cover the whole range of the latent trait. Four of the constructs (i.e., Decision Latitude; Psychological Demands; Supervisor Social Support; and Job Insecurity) contained items that were being responded at higher than the group of respondents (i.e., the items were above the group of respondents; overall: harder to indorse). On the other hand, the Coworker Social

Support construct had items that were lower than the group of respondents indicating that the items were easy to indorse. Three of the constructs had items that were representing the same area of the latent trait (i.e., items that overlap; Decision Latitude, Psychological Demands and Supervisor Social Support items). A factor analysis and a dimensionality test from Winsteps (i.e., diagnosis statistics) was completed to ensure unidimensionality and only two constructs had a high percentage of explained variance, pointing towards a more unidimensional construct. For local independence, each construct demonstrated low MNSQ's which is indicative of dependency. Although the values were not significantly low, the assumption of local independence may be compromised from the low values for each construct.

Comparison to Previous Research

Karasek and Theorell (1980) included a total of 38 items when implementing the full JCQ (i.e., Decision Latitude: Skill Discretion, 6, Decision Authority, 4; Psychological Demands 5, Job Insecurity, 3; Physical Exertion, 1; Hazardous Exposure: hazardous condition exposure, 5 and toxic exposure 3; Social Support: Supervisor Social Support,4 and Coworker Social Support 4). They conducted internal consistency measures on the items of the psychosocial work dimensions on the U.S QES, however in contrast to the current study, they compared men and women. Though their Social Support scale (which was one scale) was roughly the same for both men and women, it was still comparable to the current study. Yet they did not conduct any sort of factor analysis to measure the separate dimensions.

Karasek et al. (1998) mentioned, in their cross country comparison of the five question QES, that there was a strong factor pattern for both men and women. For the current study that was not the case. Albeit there were loadings that were stronger than others (i.e., Supervisor Social Support vs. Psychological Demands), perhaps studying the males and females as separate samples may have strengthened the factor loadings.

Karasek et al. (1998) also found that Skill Discretion item #2 (i.e., "repetitive work), Psychological Demands #5 (i.e., "conflicting demands"), and Psychological Demands #8 (i.e., "wait on others") had inconsistent loading patterns. Conversely, the current study found that "repetitive work" and "wait on others" had inconsistent loading pattern. This was noted when a hierarchical structure appeared after some items were removed. Yet, the Psychological Demands item #5 had a consistent loading pattern when it was factor analyzed.

Sale and Kerr (2002) performed the most recent examination of the psychometric properties of the JCQ (14 core items). The results from their internal consistency measures ranged from satisfactory to good. They stated that low correlations may signal overlapping items. Sale & Kerr (2002) found low internal consistencies for the Psychological Demands and Decision Authority scales, whereas the current study produced low internal consistencies for the Job Insecurity scale. A one and two factor solution for Psychological Demands was insignificant, which was also the case for the current study. Four items had to be deleted from the scale because they violated the assumption of unidimensionality.

Summary

Previous research examined the JCQ using CTT, as well as used a different number of items within each scale. Thus, there is inconsistency with previous research. The current study examined the JCQ with IRT. One interesting finding is that although the Social Support scale is theoretically made up of Coworker Social Support and Supervisor Social Support, the current study found that both the Coworker Social Support scale and the Supervisor Social Support scale were considered two separate constructs. This should red flag how social support is measured as the social support of a supervisor may be vastly different from the social support of a coworker. However, some similar results were found with regards to items that did not fit when using factor analysis.

The significance of using IRT is important within the educational field because not only is it another type of measurement, but it also provides benefits in regards to test development and examining tests that are already developed. One example is adaptive testing. Adaptive testing allows the test developer to execute a shorter test with fewer items without interfering with the reliability and validity of the test. Also, being focused on the item level makes is a useful tool for examining questionnaires such as the JCQ to accumulate validity evidence. In all, IRT is significant in both test development and examining previously developed tests, questionnaires, etc within the educational field.

Construct Validity

There are two types of construct validity; (1) construct underrepresentation and (2) construct irrelevant variance (Messick, 1989). The current study investigated the psychometric properties of the JCQ by accumulating more validity evidence. Dimensionality of the constructs had to be assessed in order to ensure that unidimensionality of each of the constructs was met, as

part of one of the assumptions of IRT (Linacre, 2006). As a secondary measure, validity was investigated to determine whether or not items fit each construct. Overall, each construct demonstrated construct underrepresentation, with the items not fully covering the breadth of the latent trait.

Advantages

There are advantages of implementing IRT over CTT methods including; (1) Allowing researchers to see what standard error of measurement is estimated at every level of the trait; (2) Each item provides information about the latent trait (Scherbaum, 2006); (3) Being able to use a set of values from items of a particular group of individuals; and (4) Estimating participant ability of only a certain group of items (Hambleton et al., 1991). As for polytomous data, IRT is a good tool for looking at strongly agree-to strongly disagree items and saying where they fit and where they do not fit within the construct. In all, IRT provides some advantages when assessing polytomous data, as well as, has advantages over CTT methods.

Limitations

Care should be taken when using IRT based programs such as Winsteps to conduct instrument development, as items must be unidimensional to fit the model. As such, many items may be deleted from a potential scale and only the items that 'fit;' the model are analyzed. In order to address the assumptions of IRT some potentially useful items may be deleted following the Factor Analysis stage and never evaluated using IRT.

A limitation can also be accounted for by the JCQ model. Problems with the JCQ (i.e., EFA) might not stem from how the questionnaire was developed but from what theory the questionnaire was derived from (i.e., DC and DCS). Kristensen (1995) has suggested that many of the previous empirical studies had many downfalls in how they explained the interaction between the psychosocial environment and psychological well-being. Thus, there may be areas not being measured such as personality factors and home stressors that also can affect work stress.

The limitations of the study include that it is a secondary data analysis (i.e., the data was not collected for the purpose of examining the psychometric properties of the JCQ), and it only incorporated four constructs and 30 items from the JCQ recommended version 1.11 (i.e., recommended contains 5 scales and 41 items). The JCQ full recommended version (i.e., 5 scales

and 49 items) was not examined (i.e., may provide a fuller picture of work stress) and some items had to be deleted because they did not meet the assumption of unidimensionality.

Opportunities for Future Research

As the current study simply analyzed the items of the JCQ recommended version 1.11, future areas of research can be more specific in the area of individual measurement error (e.g. nursing), adaptive testing (i.e., tests geared towards individual's ability), and DIF (i.e., item bias). Briggs and Wilson (2007) have introduced a new approach termed "Generalizability in Item Response Modeling (GIRM)", which includes both CTT and IRT modeling. Consequently, this could be an innovative way to overcome certain downfalls of separately applying either CTT or IRT modeling. As there are advantages and disadvantages to using both CTT and IRT, perhaps a new focus could be on using the advantages of both methods to examine a scale or questionnaire or more specifically the JCQ. However, since the results indicated that the items do not fully measure what they are suppose to measure, care should be taken when administering this questionnaire, as some areas are lacking.

Other opportunities for new research include replicating the study with an independent sample, examining the difference between males and females, and considering other polytomous models. Therefore, more research on the JCQ is required, in both accumulation of validity and examination of the items, in order to have a questionnaire that fully measures work stress.

Regardless, this study provides additional validity evidence for the use of the JCQ version 1.11 and highlights areas of psychometric weakness. As no instrument will ever be completely valid (Messick, 1989) the most a validation study such as this can hope to do is to add a unique piece of validity evidence to the existing literature.

References

Andries van der Ark, L. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*, 273–282.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, California: Brooks/Cole Publishing Company.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp.397-472). Reading, MA: Addison-Wesley.

Bongers, P.M., de Winter, C.R., Kompier, M. A. J., & Hildebrandt, V.H. (1993). Psychosocial factors at work and musculoskeletal disease. *Scandinavian Journal of Work, Environment & Health, 19*, 297-312.

Botempo, R. (1993). Translation fidelity of psychological scales. *Journal of Cross-Cultural Psychology, 24*, 149-166.

Bradley, K. D., & Sampson, S. O. (2005). Improving data collection through Rasch measurement: A continuing study of supply and demand. Paper presented at the American Educational Research Association annual meeting, Montreal, Canada.

Breaugh, J.A., & Colihan, J.P. (1994). Measuring facets of job ambiguity: Construct validity evidence. *Journal of Applied Psychology, 79*, 191-202.

Briggs, D.C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*, 133-155.

Briner, R.B. (2000). Relationships between work environments, psychological environments and psychological well-being. *Occupational Medicine, 50*, 299-303.

Brisson, C., Blanchette, C., Guimont, C., Dion, G., Moisan, J., & Vézina, M. (1998). Reliability and validity of the French version of the 18-item karasek job content Questionnaire. *Work & Stress, 12*, 322-336.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Cheng, Y., Luh, W.M., & Guo, Y.L. (2003). Reliability and validity of the Chinese version of the job content questionnaire in Taiwanese workers. *International Journal of Behavioral Medicine, 10*, 15-30.

Cronbach, L.J., & Meehl, P. E. (1995). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Dae-Yeop, H. (2002). Classical test theory and item response theory: analytical and empirical comparisons. *Proceedings at the Annual Meeting at the Southwest Educational Research Association, Austin, TX*, 1-29.

de Lange, A.H., Taris, T.W., Kompier, M.A.J., Houtman, I.L.D., & Bongers, P.M. (2003). "The very best of the millennium": longitudinal research and the demand control (-support) model. *Journal of Occupational Health Psychology, 8*, 282, 305.

Dodeem, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement*, *41*, 261-270.

Dollard, M. F. (1996). Work stress: Conceptualizations and implications for research methodology and workplace intervention. Unpublished doctoral dissertation, University of South Australia, Adelaide, South Australia, Australia.

Edimansyah, B.A., Rusli, B.N., Naing, L., & Mazalisah, M. (2006). Reliability and construct validity of the Malay version of the job content questionnaire (JCQ). *The Southeast Asian Journal of Tropical Medicine and Public Health, 37*, 412-416.

Ellis, B.E., Becker, P., & Kimmel, H.D. (1993). An item response theory evaluation of an English version of the trier personality inventory (TPI). *Journal of Cross-Cultural Psychology, 24*, 133-148.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, N.J Lawrence, Erlbaum Associates. Retrieved April 29, 2007 from http://www.netlibrary.com.cyber.usask.ca/Reader/.

Eum, K.D., Li, J., Jhun, H.J., Park, J.T., Tak, S.W., Karasek, R., et al. (2006). Psychometric properties of the Korean version of the job content questionnaire: Data from health care workers. *International Archives of Occupational and Environmental Health*, *80*, 497-504.

Evolahti, A., Hultcrantz, M., & Collins, A. (2006). Women's work stress and cortisol levels: A longitudinal study of the association between the psychosocial work environment and serum cortisol. *Journal of Psychosomatic Research, 61*, 645-652.

Frisbie, D.A (2005). Measurement 101: Some fundamentals revisited. Presidential address to the National Council on Measurement in Education. *Educational Measurement: Issues and Practice, 24*, 21-28.

Gorsuch, R. L. (1983). *Factor analysis (2nd ed.)*. Hillsdale, NJ: Erlbaum.

Haertal, E.D. (2006). Reliability. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp, 66-67). Phoenix, Arizona: American Council on Education and the Oryx Press.

Hambleton, R. K. (1993). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed, pp. 147-200). Pheonix, Arizona: American Council on Education and the Oryx Press.

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *ITEMS: The Instructional Topics in Educational Measurement Series*, 253-260. Retrieved March 20, 2007 from http://www.ncme.org/pubs/items.cfm.

Hambleton, R. K., & Swaminathan, H. (1985*). Item response theory: Principles and applications*. Boston: Kluver- Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Publications.

Harris, D. (1989). Comparison of 1-, 2-, and 3- parameter IRT models. *ITEMS: The Instructional Topics in Educational Measurement Series*, 157-163. Retrieved March 27, 2007 from http://www.ncme.org/pubs/items.cfm.

Harvey, R.J. (1999). Item response theory. *The Counseling Psychologist, 13*, 353-383.

Hensen, R.K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Houtman, I., Kornitzer, M., DeSmet, P., Koyuncu, R., DeBacker, G., Pelfrene, E., et al. (1999). Job stress, absenteeism, and coronary heart disease: European cooperative study (the JACE study). *The European Journal of Public Health, 9*, 52- 57.

Johnson, J.V., & Hall, E.M. (1988). Job strain, work place social support, and cardiovascular disease: A cross-sectional study of a random sample of the Swedish working population. *American Journal of Public Health, 78*, 1336-1342.

Karasek, R. A. (1979). Job demands, job decision latitude and mental strain: implications for job redesign. *Administrative Science Quarterly, 24*, 285-308.

Karasek, R. (1985). *Job content instrument questionnaire and user's guide, version 1.1*. Department of Industrial and Systems Engineering, University of Southern California, Los Angeles.

Karasek, R., Brisson, C., Kawakami, N., Houtman, I., & Bongers, P. (1998). The job content questionnaire (JCQ): An instrument for internationally comparative assessments of psychosocial job characteristics. *Journal of Occupational Health Psychology, 4*, 322-355.

Karasek, R., & Theorell, T. (1990). *Healthy work: stress, productivity, and the reconstruction of working life*. New York: Basic Books, Inc., Publishers.

Karasek, R., Triantis, K.P., & Chaudhry, S.S. (1982). Coworker and supervisor support as moderators of associations between tasks characteristics and mental strain. *Journal of Occupational Behavior, 3*, 181-200.

Kawakami, N., Kobayashi, F., Araki, S., Harratani, T., & Furui, H. (1995). Assessment of job stress dimensions based on the job demands- control model of employees of telecommunications and electric power companies in Japan: Reliability and validity of the Japanese version of the Job Content Questionnaire. *International Journal of Behavioral Medicine, 2*, 358-375.

Kivimäki, M., Head, J., Ferrie, J.E., Shipley, M.J., Brunner, E., Vahtera, J., et al. (2006A). Work stress, weight gain and weight loss: Evidence for bidirectional effects of job strain on body mass index in the Whitehall II study. *International Journal of Obesity, 30*, 982–987.

Kivimäki, M., Virtanen, M., Elovainio, M., Kouvonen, A., Vaananen, A., & Vahtera, J. (2006B). Work stress in the etiology of coronary heart disease: A meta-analysis. *Scandinavian Journal of Work, Environment and Health, 32*, 431- 442.

Krantz, G., & Lundberg, U. (2006). Workload, work stress, and sickness absence in Swedish male and female white-collar employees. *Scandinavian Journal of Public Health, 34*, 238-246.

Kristensen, T.S. (1995). The demand-control-support model: Methodological challenges for future research. *Stress Medicine, 11*, 17-26.

Landsbergis, P.A., Schnall, P.L., Warren, K., Schwartz, J.E., & Pickering, T. (1994). Association between ambulatory blood pressure and alternative formulations of job strain. *Scandinavian journal of work, environment health, 20*, 349-363.

Landsbergis, P.A., & Theorell, T. (2000). Measurement of psychosocial workplace exposure variables. *Occupational Medicine, 15*, 163-188.

Linacre, J. M. (2006) *A User's Guide to Winsteps Rasch-Model Computer Programs*. Chicago, IL: MESA Press.

Lohr, S. (1996). Though upbeat on the economy, people still fear for their jobs. Retrieved on April 23, 2007 from the New York Times Web Site http://www.nytimes.com/specials/downsize/1229econ-jobs-uncertain.html.

Lord, F.M. (1953). The relation of test score to the trait underlying the test source. *Educational and psychological measurement 13*, 517-549.

Lord, F.M., & Novick, M.R. (1952). *A theory of test scores (Psychometric Monograph N o. 7)*. Iowa City, IA: Psychometric society.

Lord, F.M., & Norvick, M.R. (1968). *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley Pub Co.

Loyd, B. (1988). *Implications of item response theory for the measurement practitioner*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (Vol.3, pp. 13-103). New York: American Council on Education/ Macmillan.

McCarty, F. (2005). Item response theory. In D. Iorio & C. Konicki (Eds.), (1st) *Measurement in health behavior: methods for research and education*. San Francisco: Jossey-Bass.

Munce, S.E.P., Weller, I., Blackmore, E.K.R, Heinmaac, M., Katz, & Stewart, D.E. (2006). The role of work stress as a moderating variable in the chronic pain and depression association. *Journal of Psychosomatic Research, 61*, 653- 660.

Norman, G.R., & Streiner, D.L. (2003). Pretty darned quick (3rd ed.). Hamilton, Ontario: BC Decker Inc.

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric theory (3rd ed.). McGraw Hill, New York.

Ostini, R., & Nering, M.L. (2006). *Polytomous item response theory models*. Thousand Oaks, California: Sage Publications.

Reise, S.P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Rousseau, D.M. (1995). *Psychological contracts in organizations*. Thousand Oaks, California: Sage Publications.

Sampson, S. O., & Bradley, K. D. (2005). *Quality control in survey design: Evaluating a rating scale of educators' attitudes toward differentiated compensation.* Paper presented at the American Educational Research Association annual meeting, Montreal, Canada.

Sanne, B., Torp, S., Mykletun, A., & Dahl, A.A. (2005). The Swedish demand-control- support questionnaire (DCSQ): Factor structure, and internal consistency in a large population. *Scandinavian Journal of Public Health, 33*, 166-174.

Santavirta, N. (2003). Construct validity and reliability of the Finnish version of the demand-control questionnaire in two samples of 1028 teachers and 630 nurses. *Educational Psychology, 23*, 423-437.

Sauter, S.L., Hurrell Jr., J.J., & Cooper, C.L. (Eds.) (1989). *Job control and worker health*. West Sussex, England: John Wiley & Sons Ltd.

Scherbaum, C. A. (2006). *Applications of item response theory to measurement issues in leadership research*. Greenwich, Conn.: JAI Press.

Schreurs, P. J.G., & Taris, W.T. (1998). Construct validity of the demand-control model: A double cross-validation approach. *Work & Stress, 12*, 66-84.

Shaw, J.B., & Weekley, J.A. (1985). The effects of objective work-load variations of psychological strain and post-work-load performance. *Journal of Management, 11*, 87-98.

Sorensen, H.T., Sabroe, S, & Olsen, J. (1995). A framework for evaluation of secondary data sources for epidemiological research. *International Journal of Epidemiology, 25*, 435-442.

SPSS for Windows, Rel. 15.0.1. 2006. Chicago: SPSS Inc.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, New Jersey: Lawrance Erlbaum Associates.

Statistics Canada (1996). *National Population Health Survey: Public use micro-data documentation*. Ottawa: Author.

Storms, G., Casaer, S., De Wit, R., Vandenbergh, O., & Moens, G. (2001). A psychometric

evaluation of the Dutch version of the job content questionnaire and of a short direct questioning procedure. *Work & Stress, 15*, 131-143.

Streiner, D.L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99-103.

Tejada, A.J.R., & Rojas, O.M.L. (2005). Application of an IRT polytomous model for measuring health related quality of life. *Social Indicators Research, 74*, 369-394.

Thissen, D., & Wainer, H. (2001). *Test scoring.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Theorell, T. (1996). The demand-control-support model for studying health in relation to the work environment: An interactive environment. In K. Orth-Gomér & N. Schneiderman (Eds.), *Behavioral approaches to cardiovascular disease prevention (*pp.69-85). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Thurstone, L.L. (1947). Multiple-factor analysis: A development and expansion of the vectors of mind. Chicago, Illinois: University of Chicago Press.

Traub, R.E., & Rowley, G.L. (1991). Understanding reliability. *ITEMS: The instructional topics in educational measurement series, 171-179*. Retrieved on April 15, 2007 from http://www.ncme.org/pubs/items.cfm.

Van der Doeuf, M., & Maes, S. (1999). The job demand-control (-support) model and psychological well-being: A review of 20 years of empirical research. *Work & Stress, 13*, 87-114.

Warr, P. (1994). A conceptual framework for the study of work and mental health. *Work & Stress, 8*, 84-97.

Weiss, D.J. (1983). *New horizons in testing*. New York, New York: Academic Press, Inc.

Wright, B.D. (1977). Solving measurement problems with the rasch model. Retrieved January 24, 2008 from http://rasch.org/memo42.htm.

Wright B. D. (1996) Reliability and separation. *Rasch Measurement transactions 9*, 472.

Zenisky, A.L., Hambleton, R.K., & Sireci, S.G. (2001). Effects of local item dependence on the validity of IRT item, test and ability statistics. *Association of American Medical Colleges, Washington, DC*. Retrieved on August 12, 2007 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/19/db/01.pdf.

Appendix A

Table A1

Comparing the Scales and Numbers of Questions in the Full Recommended JCQ and the "Core QES"

Scale	Core QES JCQ	Full Recommended JCQ
1. Decision Latitude		
a. Skill discretion	6	6
b. Decision authority	3	3
c. Skill underutilization	2[b]	2[b]
d. Work group decision authority (new)		3
e. Formal authority (new)		2
f. Union/ representative influence (new)		3
2. Psychological demands and mental workload		
a. General psychological demands	4	5
b. Role ambiguity	1	1
c. Concentration (new)		1
d. Mental work disruption (new)		2
3. Social Support		
a. Socioemotional (coworker)	2	2
b. Instrumental (coworker)	2	2
c. Socioemotional (supervisor)	2	2
d. Instrumental (supervisor)	2	3

Table A1 continued.

Comparing the Scales and Numbers of Questions in the Full Recommended JCQ and the "Core QES"

Scale	Core QES JCQ	Full Recommended JCQ
e. Hostility (coworker) (new)		1
f. Hostility (supervisor) (new)		1
4. Physical Demands		
a. General physical loading	1	1
b. Isometric load (new)		2
c. Aerobic load (new)		2
5. Job insecurity		
a. General job insecurity	3	4
b. Skill obsolescence (new)		2
Total Questions	27	49

Note. JCQ = Job Content Questionnaire; QES = Quality of Employment Surveys.

[a] Eight new scales/dimensions and additional items were added to make the Recommended JCQ format. [b] Education was also used in this scale.

Adopted from Karasek et al. (1998)

Appendix B

Table B1

Psychometric properties of the English version of the JCQ

Source	Version	Items	Sample	Validity estimates	Reliability estimates	Theory
Karasek (1979)	English	Total:19 items From U.S. Decision latitude: skill discretion (4) and decision authority (4) Job Demands (7) From Sweden - Decision Latitude: intellectual discretion (2) and expert rating of skill level required - Job Demands (2)	Random sample of a full adult population (Swedish Survey; 1:1000)	- ANOVA	none	CTT
Karasek & Theorell (1980)	English	Total: 38 items - Decision Latitude: skill discretion (6) and decision authority (4) -Psychological Demands (5) -Job Insecurity (3) -Physical Exertion (1) -Hazardous Exposure: hazardous condition exposure (5) and toxic exposure (3) -Social Support: supervisor social support (4) and coworker social support (4)	Total: 4,503 Random sample	- Correlation matrix	- Internal Consistency Cronbach's alpha - Cross- survey correlation: test re-test	CTT

Table B1 continued.

Psychometric properties of the English version of the JCQ

Source	Version	Items	Sample	Validity estimates	Reliability estimates	Theory
Karasek, Brisson, Kawakami, Houtman, & Bongers (1998)	English	Less than the Full recommended-Less than version 1.11 Final questionnaire used: The five QES version	Total: 16, 601 Comparative study Canada-Quebec United States Japan Netherlands	- ANOVA - Concurrent validity: correlations between scales and subscales - Factor validity analysis	- Internal Consistency: Cronbach's alpha	CTT
Sale & Kerr (2002)	English	Total: 14 items Decision latitude: skill discretion (6) and decision authority (3) Psychological demands (5)	Total: 900 employees Random sample of hospital staff	- CFA - Goodness-to fit Index - Non-normed fit index -Comparative fit index - Incremental fit index - ANOVA	- Internal consistency: Cronbach's alpha - Inter-correlations: Pearson correlations - Item-total correlations	CTT

Appendix C

Table C1

A Summary: Psychometric Properties of several versions of the JCQ

Source	Version	Items	Sample	Validity estimates	Reliability estimates	Theory
Edimansyah, Rusli, Naing, & Mazalisah (2006)	Malay	Total: 21 items Decision Latitude (8) Psychological demand (7) Social Support (6)	Total: 50 Mostly male (90%) sample automotive assembly plant workers	- Construct Validity: EFA	- Internal Consistency: Cronbach's alpha	CTT
Santivirta (2003)	Finnish	Total: 11 items Psychological Demands (5) Decision Latitude (6): skill discretion (4) and decision latitude (2)	First Sample: Primary, Secondary and High School Teachers (1028) Sample 2: Registered Nurses (603)	- EFA - CFA	- Internal Consistency: Cronbach's alpha	CTT
Cheng, Luh, & Guo (2003)	Chinese	Total: 22 items Decision Latitude: skill discretion (6) and decision authority (3) Psychological Demands (5) Social Support: support from supervisor (4) Support from coworkers (4)	Total: 1199 workers from offices and plants from four private factories participated	- Principle component analysis	- Internal consistency: Cronbach's alpha - Test re-test: Pearson correlations	CTT

Table C1 continued.

A Summary: Psychometric Properties of several versions of the JCQ

Source	Version	Items	Sample	Validity estimates	Reliability estimates	Theory
Brisson, Blanchette, Guimont, Dion, & Vézina (1998)	French	Total: 18 items Decision Latitude: skill discretion (6) and decision authority (3) Psychological Demands (9)	Total of 8263; half were women White collar workers employed in 20 different public and private organizations	- Factorial Validity: EFA	- Inter-correlations: Pearson Correlation Coefficient - Internal Consistency: Cronbach's alpha	CTT
Storms, Casaer, De – Wit, Vandenbergh, & Moens (2001)	Dutch	Total: 43 items An adapted JCQ	Total: 3638 workers 1995 men 1643 women White collar workers	- Factorial Validity: EFA	- Spearmann Brown Formula - Split half correlation - Pearson Product moment correlation	CTT
Kawakami, Kobayashi, Araki, Haratani, & Furiu (1995)	Japanese	Total: 22 items Decision Latitude Psychological Demand Supervisor Support Coworker support	Total: 472 men 108 women Clerical workers from a Telecommunication company	- Principle component factor analysis - EFA - ANCOVA - Multiple Linear Regression	- Pearson correlation - Internal consistency: Cronbach's alpha - Spearman's Rank	CTT

Table C1 continued.

A Summary: Psychometric Properties of several versions of the JCQ

Source	Version	Items	Sample	Validity estimates	Reliability estimates	Theory
Eum, Li, Jhun, Park, Tak, Karasek & Cho (2006)	Korean	Total: 49 items Decision Latitude (9) Psychological demand (5) Social support (8) Macro-level decision latitude (6) Job insecurity (3) Physical exertion (1)	Total: 290 females 48 males Nurses, technicians, administrative personnel and employees in the nutrition department	- Factor Validity: EFA - Criterion validity: Multiple regression	- Internal Consistency: Cronbach's alpha - Test re-test reliability: Pearson correlations	CTT
Sanne, Torp, Mykletunm & Dahl (2005)	Swedish	Total: 12 items Decision latitude (6): skill discretion (4) and decision authority (2) Social support (6)	Total: 29,400: not sure what proportion of the study was employed	- Principle component analysis	- Inter-correlations: Pearson correlation - Internal consistency: Cronbach's alpha	CTT

Appendix D

Table D1

Frequencies of the sample from Janzen's (2006) study (n = 1160).

Variable		Frequency	%	Mean	St. D.
Gender	Male	486	41.9		
	Female	674	58.1		
Age				36.03	0.214
Marital Status	Married	641	55.3		
	Living with a partner	138	11.9		
	Widowed	20	1.7		
	Separated	96	8.3		
	Divorced	142	12.2		
	Single	123	10.6		
Type of Occupation	Management and professional	200	17.2		
	Teaching and related	171	14.7		
	Medical and Health	178	15.3		
	Clerical/Sales/Service	450	38.8		
	Construction trades	52	4.5		
	Transportation	26	2.2		
	Farmer	2	0.2		
	Self-employed	39	3.4		
	Civil Servant	26	2.2		

Table D1 continued.

Frequencies of the sample from Janzen's (2006) study (n = 1160).

Variable	Frequency	%	Mean	St. D.
No Response	16	1.4		
Amount of hours spent at work			39.6	11.3

Appendix E

Job Strain Questionnaire

The next set of questions asks you to think about different aspects of your job. If you have more than one job, please consider how each question applies to your main job.

a) Please tell me whether you strongly disagree, disagree, agree or strongly agree with each of the following statements. (Interviewer: Please use the following key to indicate the participant's response. The numbers are not intended to be read out to the participant)

1	**2**	**3**	**4**
Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**

Subscale	**Item**	**Number**
Skill Discretion	My job requires that I learn new things.	1 2 3 4
Skill Discretion	My job requires a lot of repetitive work.	1 2 3 4
Skill Discretion	My job requires me to be creative.	1 2 3 4
Skill Discretion	My job requires a high level of skill.	1 2 3 4

Skill Discretion	I get to do a variety of different things on my job.	1 2 3 4
Skill Discretion	I have the opportunity to develop my own special abilities.	1 2 3 4
Decision Authority	My job allows me to make a lot of decisions on my own.	1 2 3 4
Decision Authority	On my job, I have very little freedom to decide how I do my work.	1 2 3 4
Decision Authority	I have a lot to say about what happens on my job.	1 2 3 4
Psychological Demands	My job requires working very fast.	1 2 3 4
Psychological Demands	My job requires working very hard.	1 2 3 4
Psychological Demands	I am not asked to do much work.	1 2 3 4
Psychological Demands	I have enough time to get the	1 2 3 4

	job done.	
Psychological Demands	The demands that other people make of the often conflict.	1 2 3 4
Psychological Demands	My job requires long periods of intense concentration on the task.	1 2 3 4
Psychological Demands	My job is very hectic.	1 2 3 4
Psychological Demands	Waiting on work from other people or departments often slows me down on my job.	1 2 3 4
Psychological Demands	My tasks are often interrupted before I can finish them so that I have to go back to them later.	1 2 3 4
Coworker Social Support	People I work with are competent in doing their jobs.	1 2 3 4
Coworker Social Support	People I work with take a personal interest in me.	1 2 3 4

Coworker Social Support	People I work with are friendly.	1 2 3 4
Coworker Social Support	People I work with are helpful in getting the job done.	1 2 3 4
Job Insecurity	My job security is good.	1 2 3 4
Job Insecurity	My prospects for career development and promotions are good.	1 2 3 4
Job Insecurity	In five years, my skills will still be valuable.	1 2 3 4
Supervisor Social Support	My supervisor is concerned about the welfare of those under him/her.	1 2 3 4
Supervisor Social Support	My supervisor pays attention to what I am saying.	1 2 3 4
Supervisor	My supervisor is helpful in getting the job done.	1 2 3 4

Social Support		
Supervisor Social Support	My supervisor is successful in getting people to work together.	1 2 3 4
Supervisor Social Support	I am exposed to hostility or conflict from my supervisor.	1 2 3 4

*Janzen (2006) uses a fifth response category (i.e., 5: Unsure)

Appendix F

Table F1

Item polarity for construct #4 (supervisor social support)

Item	Raw Score	Count	Measure	Model S.E.	Infit		Outfit		Point measure correlation
					MNSQ	ZSTD	MNSQ	ZSTD	
Supervisor Social Support: (4) My supervisor is good at getting people to work together.	1729	658	57.08	0.64	1.46	7.1	1.43	4.5	0.72
Supervisor Social Support: (1) My supervisor is concerned about the welfare of those under him/her.	1958	658	47.21	0.68	0.94	-1.1	0.85	-1.7	0.78
Supervisor Social Support: (3) My supervisor is helpful in getting the job done.	1913	658	49.24	0.67	0.88	-2.1	0.82	-2	0.79
Supervisor Social Support: (2) My supervisor pays attention to what I am saying.	1974	658	46.47	0.68	0.7	-5.5	0.64	-4.5	0.81

Appendix G

```
INPUT: 1160 Persons  4 Items  MEASURED: 690 Persons  4 Items  5 CATS
3.63.2
-------------------------------------------------------------------------
---

OBSERVED AVERAGE MEASURES FOR Persons (unscored) (BY OBSERVED CATEGORY)
-10   0    10    20    30    40    50    60    70    80    90
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   Item
|        1        2        3                4               |    4  SSS4
|                                                          |
|                                                          |
|  1        5  2         3                  4               |    3  SSS3
|    1        2         3                  4               |    1  SSS1
| 1           2         3 5                 4               |    2  SSS2
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   Item
-10   0    10    20    30    40    50    60    70    80    90

                         1
31    2  2  4  3  7  7   4    6    4     4                 6
27    5  3  4  6  6  0   0    5    9     7                 6  Persons
         S               M               S             T
```

Figure G1. Empirical item measures for construct #4 (supervisor social support)

Appendix H

Table H1

Dimensionality for construct #1 (decision latitude), construct #2 (psychological demands), construct #3 (coworker social support), construct #4 (supervisor social support) and construct #5 (job insecurity).

Construct	Total Variance in observations	Variance explained by measures	Unexplained variance
1	100.0 %	68.3%	31.7%
2	100.0%	77.4%	22.6%
3	100.0%	55.7%	44.3%
4	100.0%	85.8%	14.2%
5	100.0%	85.7%	14.3%

Appendix I

Table I1

Item misfit for construct #4 (supervisor social support)

Entry #	Raw Score	Count	Measure	Model S.E.	Infit		Outfit		PTMEA Corr.	Exact Obs%	Match Exp%	Item
					MNSQ	ZSTD	MNSQ	ZSTD				
4	1729	658	57.08	0.64	1.46	7.1	1.43	4.5	A72	65.2	60.7	SSS4
1	1958	658	47.21	0.68	0.94	-1.1	0.85	-1.7	B.78	71.6	64.6	SSS1
3	1913	658	49.24	0.67	0.88	-2.1	0.82	-.20	B 0.79	65.8	63.9	SSS3
2	1974	658	46.47	0.68	0.70	-5.5	0.64	-4.5	A 0.81	74.6	64.7	SSS2
Mean	1893	658.0	50.00	0.67	0.99	-0.4	0.94	-1.0		69.3	63.4	
S.D.	97.6	0.00	4.21	0.02	0.28	4.6	0.29	3.3		4.0	1.6	

Appendix J

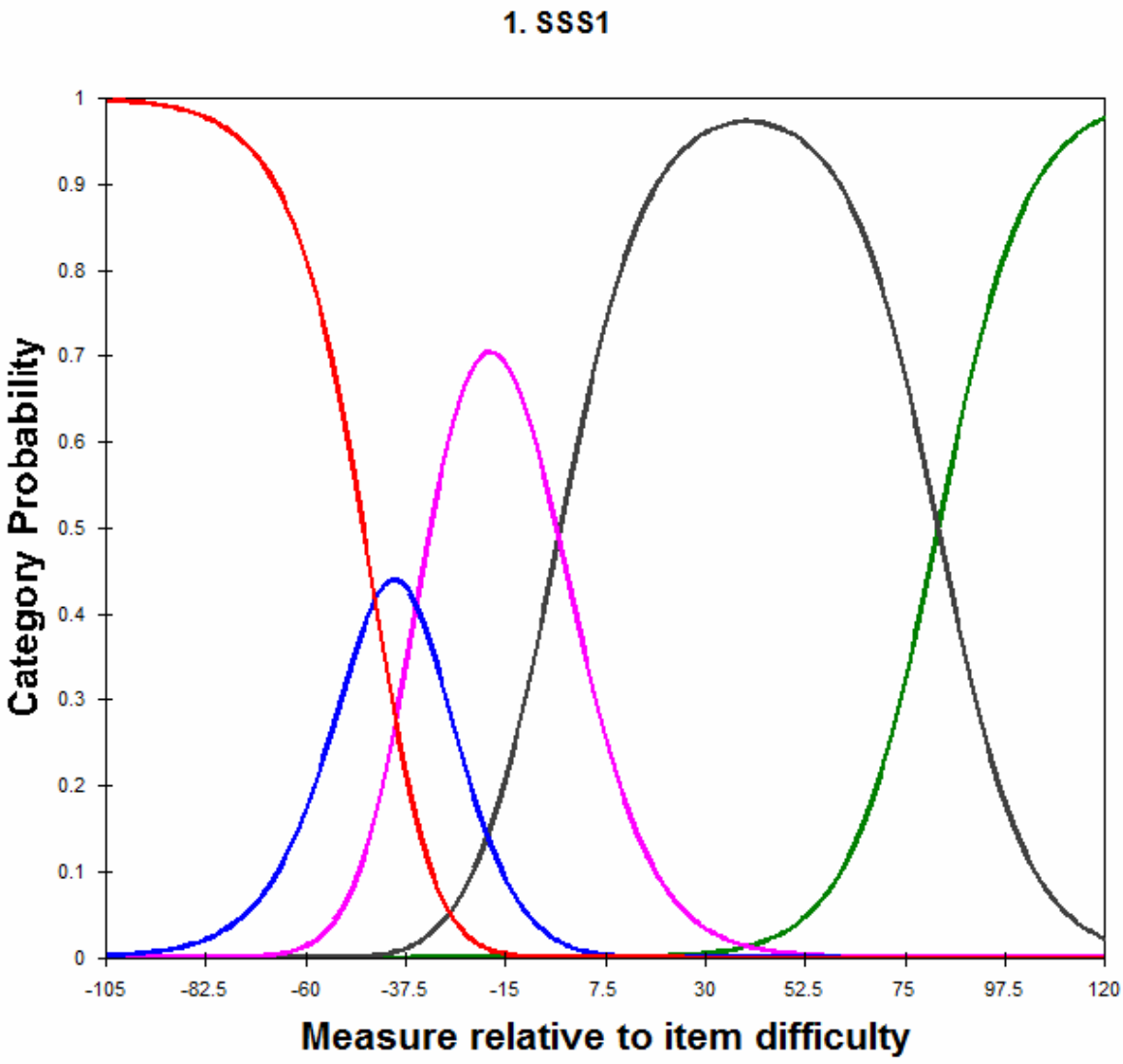
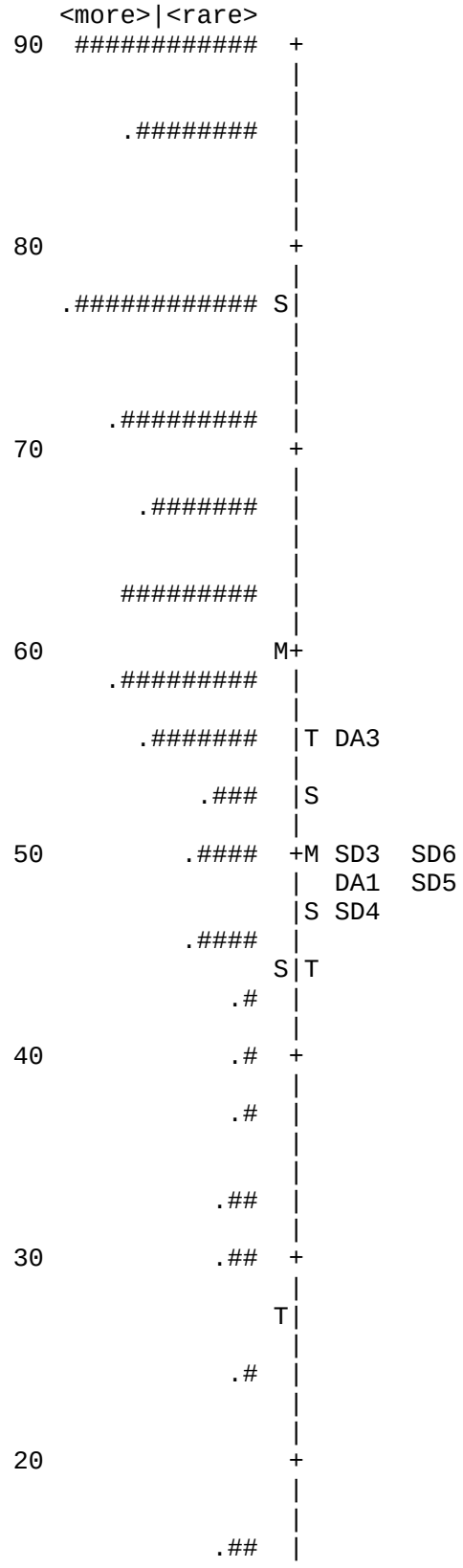


Figure J1. Category probability curve for construct #4.

Appendix K


<more>|<rare>
90
80
70
60
50
40
30
20
S
M
T DA3
S
M SD3 SD6
DA1 SD5
S SD4
S|T
T

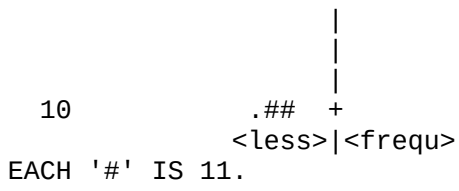


Figure K1. Item map for construct #1 (decision latitude)

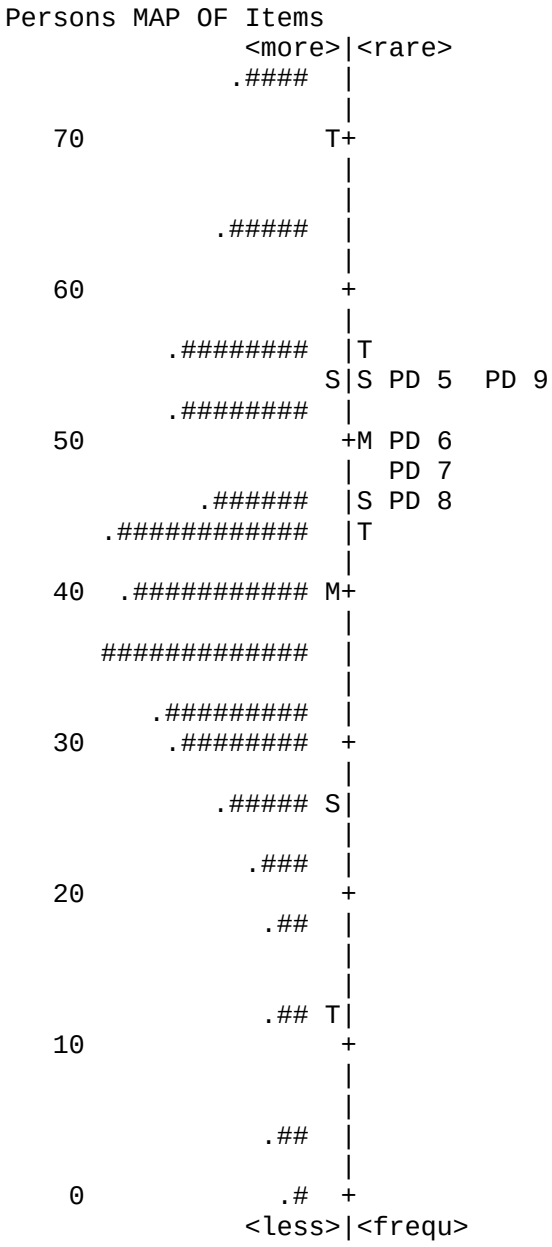
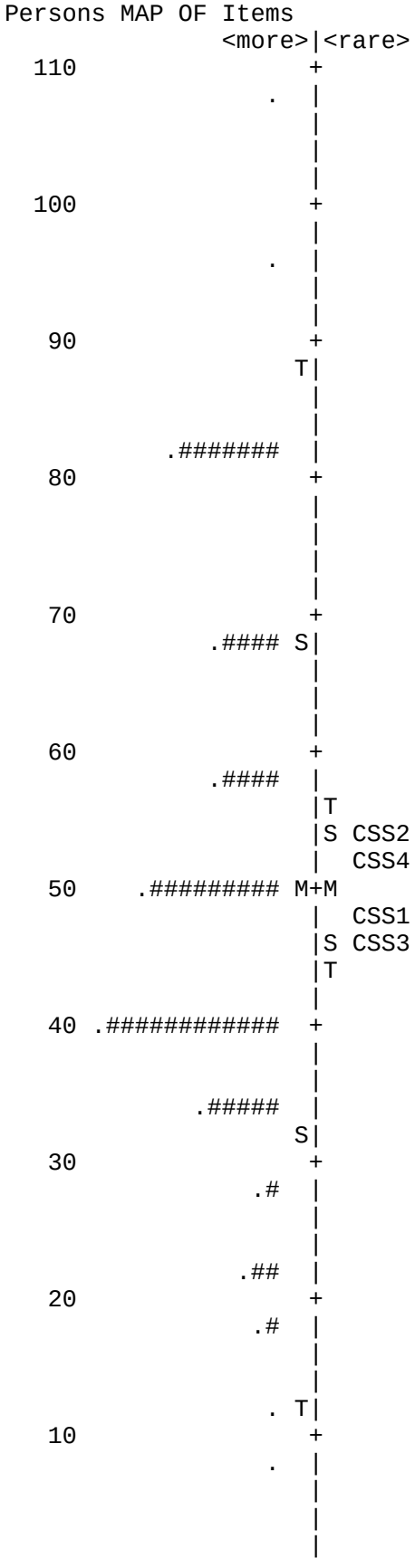


Figure K2. Item map for construct #2 (psychological demands)


Persons MAP OF Items
<more>|<rare>
110
100
90
80
70
60
50
40
30
20
10
T
S CSS2
CSS4
M+M
CSS1
S CSS3
T

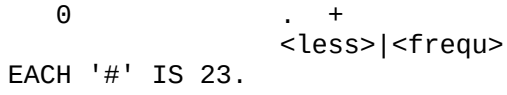


Figure K3. Item map for construct #3 (coworker social support)

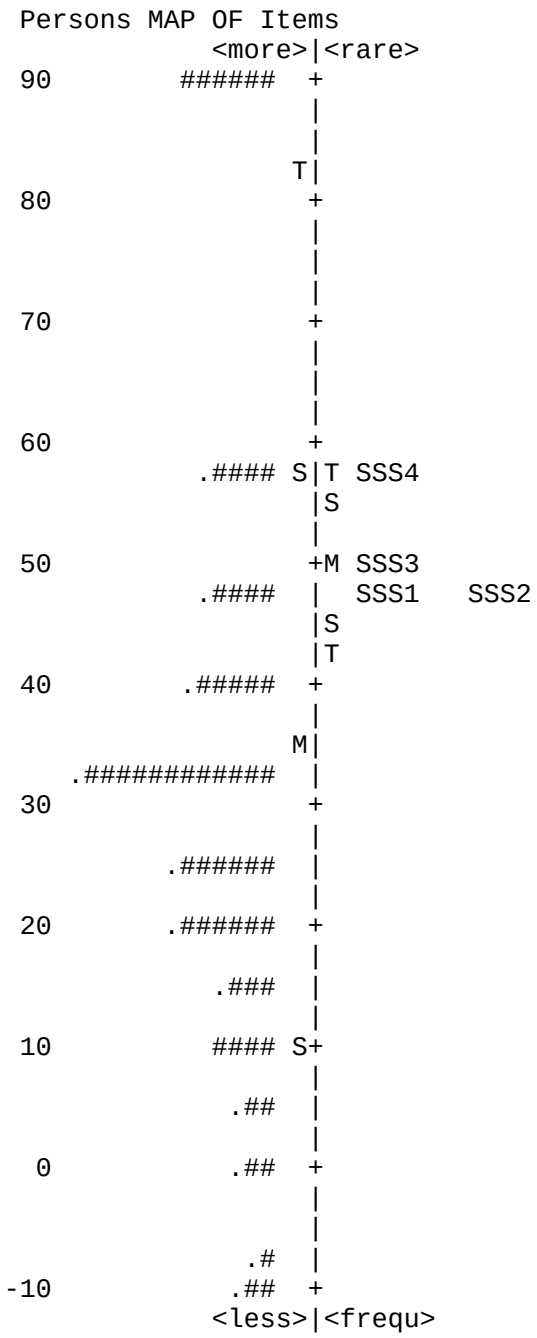


Figure K4. Item map for construct #4 (supervisor social support)

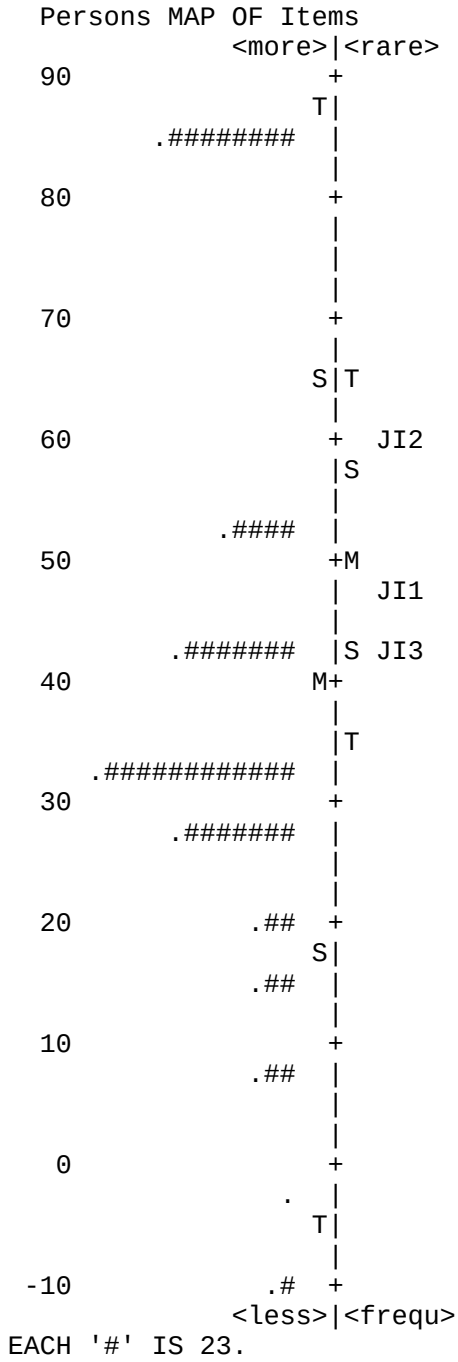


Figure K5. Item map for construct #5 (job insecurity)

Appendix L

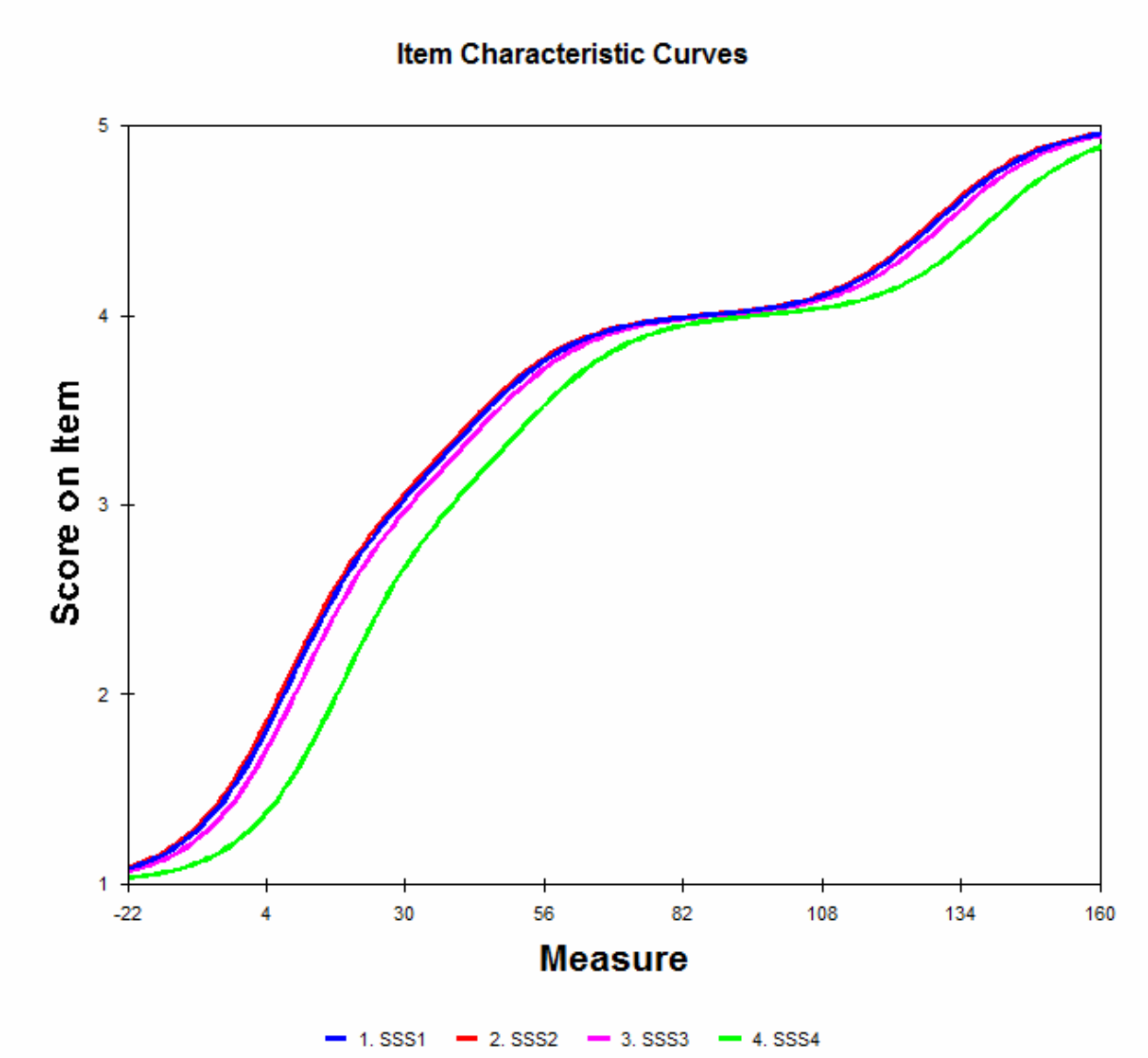


Figure L1. Construct #4: multiple ICC of 4 items.

[i] The Rasch model is similar to the 1-parameter logistic model, as it only contains one parameter (i.e., difficulty). The difference between the Rasch and IRT model is that the IRT model allows for additional parameters (i.e., guessing and discrimination), whereas the Rasch model does not (Wright, 1977).

[ii] Principal components extraction, Cattell's scree plot, and image factoring extraction were performed on the 30 items. The tables can be provided by getting in touch with the author.